IPC ACCURACY STUDY:
ANALYZING THE INTERNAL CONSISTENCY
OF IPC AFI AND AMN ANALYSES

May 2024
IPC ACCURACY STUDY:

ANALYZING THE INTERNAL CONSISTENCY OF IPC AFI AND AMN ANALYSES

Prepared by:

Erin Lentz, Associate Professor, Lyndon B. Johnson School of Public Affairs, University of Texas, Austin
Kathy Baylis, Professor and Vice-Chair, Department of Geography, University of California, Santa Barbara
Hope Michelson, Associate Professor, Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign
Chungmann Kim, PhD candidate, Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign

May 2024
**Abstract:**

The Integrated Food Security Phase Classification is the global point of reference for classification of food crises and famine and has been used in over 30 countries over the past fifteen years. It is unique. In 2022, the IPC commissioned, for the first time, an external study of the accuracy of its acute food insecurity and acute malnutrition phase classifications and population estimates. Because IPC outcomes and the underlying indicators that are used to generate IPC outcomes are latent constructs, there is no true measure of ‘accuracy.’ Thus, the first task of this study was to determine a methodology suitable to this challenge. In this study, the analysis examines IPC outcomes through the lens of internal consistency. Our approach is informed by a review of the methodological literature, the data available and findings from key informant interviews.

Drawing on multiple methods ranging from descriptive statistics to regression analyses, we compare IPC outcomes against a series of reference points informed by the IPC Technical Manual V3.1. We have rich data for acute food insecurity, and results focus on these findings. We show proof of concept findings for acute malnutrition.

To examine consistency of acute food insecurity analysis, we ask and answer four main questions.

- First, we ask what the underlying food security indicators suggest to Technical Working Groups about phase classification and population outcomes. We find that food security indicators are not highly concordant, pointing to the importance of the IPC consensus process.

- Second, we ask how Technical Working Group consensus processes use available information. We find that classifications of severity are highly consistent with the technical guidance provided in IPC Technical Manual V3.1 which serves as our benchmark. We also find that Technical Working Groups, on average, are under classifying populations in food crisis or above relative to our reference points; that is, on average they appear to be identifying fewer people as urgently food insecure (or worse) compared to what our reference analysis would point to.

- Third, we ask what could be influencing consensus outcomes. We find little evidence that any specific indicators are systematically disregarded or ignored. Consensus appears idiosyncratic: the treatment of food security indicators varies by several factors, raising questions about the comparability of the consensus outcomes. We also find that that Technical Working Groups tend to classify fewer people as food insecure when underlying food security indicators are noisier.

- Fourth, we ask how acute food insecurity projections relate to later analysis of the current status. We do not observe evidence of systematic over or under-prediction of the projections relative to the realized current study analyses.

We close with short and long-term recommendations, including reviewing food security indicator thresholds in the technical guidance, investigating pairing automated analyses with documentation to support consensus, comparability, and transparency, and undertaking a follow-up accuracy study in five years.
## CONTENTS

Acknowledgements 9
Abbreviations and Acronyms 10
EXECUTIVE SUMMARY 11

### CHAPTER 1

**INTRODUCTION** 14
1.1. Objective and approach: the role of consensus 14
1.2. Roadmap 16

### CHAPTER 2

**NATURE OF THE PROBLEM AND METHODS** 17
2.1. IPC and food security: Nature of the problem of analyzing accuracy 17
2.2. Evaluating metrics when there is no objective, measured ‘truth’ 17
   2.2.1 Qualitative findings 18
   2.2.2 Defining study 20
   2.2.3 Study parameters 21
   2.2.4 Note on language 21
2.3. Conceptual framework to understanding consensus 21
   2.3.1 A simple model of the consensus process 23
2.4. Application of methods for current status AFI 23
   2.4.1 Arithmetic (simple) mean 23
   2.4.2 Estimating weights 24
   2.4.3 Distributional analysis: “bunching” 25
   2.4.4 Residual analysis 25
   2.4.5 Noisy input data 25
2.5 Methods for projected AFI 26

### CHAPTER 3

**AFI DATA AND SAMPLES** 27
3.1. Data collection 27
3.2 Descriptive statistics – current status AFI 27
   3.2.1 Sample size, FSI availability and FSI concordance 27
   3.2.2 Reliability score 31
   3.2.3 Consensus-based phase classification outcome compared to FSI implied phase classification outcome 31

### CHAPTER 4

**AFI RESULTS** 35
4.1 Results for current status AFI 35
   4.1.1 Arithmetic (simple) mean classification estimates compared to consensus-based classifications 38
   4.1.2 Arithmetic (simple) mean population estimates (in percentages) compared to consensus-based population (in percentages) 38
   4.1.3 Logit weights across data availability and severity 41
   4.1.4 OLS Weights on population estimates 43
   4.1.5 OLS weights across TWG analysis rounds within countries 44
ACKNOWLEDGEMENTS

We thank the many users of the IPC and members of the IPC community who provided valuable insights on the IPC process, on local contexts, on food security indicators, and on measuring accuracy. We presented preliminary findings to members of the IPC TAG, Nutrition and Food Security Working Groups in February 2023 in Rome, Italy and presented our accuracy study to the IPC TAG September 2023, and to the IPC Global Steering Committee and IPC Resource Partners in November 2023. We are grateful for their insights, which greatly enriched this research.

We especially thank members of the IPC GSU and regional offices for simultaneously prioritizing our autonomy while also willingly and openly answering our (many) questions. José Lopez commissioned this study and ensured our independence. Thomas Baar, Sophie Chotard, Scott Drimie, Barbara Frattaruolo, Douglas Jayasekaran, Mariana Mizutani, Melanie Rabier, and Kamau Wanjohi ensured we had access to and understanding of all available data and IPC processes. In particular, without Kaija Korpi-Salmela and Jannie Armstrong this project would not have been possible. Any remaining errors are ours alone.
# Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFI</td>
<td>Acute food insecurity</td>
</tr>
<tr>
<td>AMN</td>
<td>Acute malnutrition</td>
</tr>
<tr>
<td>API</td>
<td>Application programming interface</td>
</tr>
<tr>
<td>CARI</td>
<td>Consolidated approach to reporting indicators of food security</td>
</tr>
<tr>
<td>CDT</td>
<td>Consolidated data tool</td>
</tr>
<tr>
<td>FCS</td>
<td>Food consumption score</td>
</tr>
<tr>
<td>FEWS NET</td>
<td>Famine early warning system network</td>
</tr>
<tr>
<td>FIES</td>
<td>Food insecurity experience scale</td>
</tr>
<tr>
<td>FSNMS</td>
<td>Food security and nutrition monitoring system</td>
</tr>
<tr>
<td>FSI</td>
<td>Food security indicator (also known as first-level outcomes in IPC process)</td>
</tr>
<tr>
<td>GAM</td>
<td>Global acute malnutrition</td>
</tr>
<tr>
<td>GSU</td>
<td>IPC's Global support unit</td>
</tr>
<tr>
<td>HDDS</td>
<td>Household dietary diversity score</td>
</tr>
<tr>
<td>HFA</td>
<td>Humanitarian food assistance</td>
</tr>
<tr>
<td>HHS</td>
<td>Household hunger scale</td>
</tr>
<tr>
<td>IPC</td>
<td>Integrated Food Security Phase Classification</td>
</tr>
<tr>
<td>IPC-API</td>
<td>Integrated Food Security Phase Classification-Application Programming Interface</td>
</tr>
<tr>
<td>LCS</td>
<td>Livelihood coping strategies</td>
</tr>
<tr>
<td>MUAC</td>
<td>Mid-upper arm circumference</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary least squares regression</td>
</tr>
<tr>
<td>rCSI</td>
<td>Reduced coping strategies index</td>
</tr>
<tr>
<td>TWG</td>
<td>Technical Working Group</td>
</tr>
<tr>
<td>WHZ</td>
<td>Weight for height z-score</td>
</tr>
</tbody>
</table>
EXECUTIVE SUMMARY

During the past fifteen years, the IPC has gained scope and prominence in global food security analysis and humanitarian allocation decisions. As the global standard for food security and nutrition analysis at national and sub-national levels, the Integrated Food Security Phase Classification (IPC) has been introduced in 35 countries. IPC analyses have guided strategic allocation of approximately six billion dollars of humanitarian food assistance per year (IPC 2023).

The accuracy and integrity of IPC evaluations are therefore of utmost importance for international and national humanitarian work. This report examines the accuracy of the IPC’s acute food insecurity (AFI) current analyses and projections and presents a proof-of-concept for doing the same for acute malnutrition (AMN).

Determining the accuracy of IPC analyses is challenging because there is no single ‘true’ measure of food insecurity against which we can compare IPC analyses. In recognition of this, we conducted qualitative research on stakeholder perceptions of IPC accuracy and reviewed methodological literature to arrive at an approach that examines the internal consistency of the Technical Working Group (TWG) consensus process, by comparing consensus-based outcomes against a series reference points and against our benchmark, the IPC V3.1 Manual Reference Table. We focus on consensus-based phase classifications and populations in IPC Phase 3 or above (crisis or worse). For methodologically interested readers, we present our approach and results in sections §2 - 5. Readers interested in the main findings will find the takeaway points in the Discussion Section §6. We present recommendations in section §7.

Based on the data available, our AFI primary sample includes 1881 TWG analyses of IPC analysis areas from 15 countries and includes three countries with multiple rounds of IPC analyses carried out by TWGs during late 2020-2022. Due to data limitations, a similar exercise could not be conducted for acute malnutrition analyses. We show proof-of-concept results for acute malnutrition population and phase classifications, recognizing that the IPC is investing in improving these data. For AFI analyses, we use underlying food security indicators to study phase classification and population outcomes. Our main current status AFI and AFI projection analyses findings are organized by four questions, presented by increasing analytical complexity for current status, followed by projections:

- First, we ask "What do the underlying indicators suggest to TWGs about phase classification and population outcomes?" This question seeks to explore the information TWGs have.
- Second, we ask "What is the consensus process doing?" This allows us to examine if population or phase classification outcomes are systematically biased.
- Third, we ask "What could be influencing consensus outcomes?" This focuses on what could be shaping the outcomes we observe.
- Fourth, we ask "What is the relationship between projections and current status classifications?" This allows us to examine if projections are systematically higher or lower than consensus classifications.

AFI current status findings include:

First, "What do the underlying indicators suggest to TWGs about phase classification and population outcomes?"
We find:

- The underlying food security indicator data vary substantially between TWG-led analysis exercises, even within a single country from one exercise to the next. The set of available food security indicators (FSIs) describe different aspects of a complex phenomenon. This renders correlations of the shares of population in crisis or worse based on individual FSI variable over space and time, underscoring the need for the contextual expertise of TWGs.
- Underlying FSI data tend to suggest different phases, implying the appropriate phase classification is not obvious. We find that TWGs’ processes – that we will refer to as “consensus-based phase classifications” – are highly consistent with our benchmark, the technical guidelines laid out in the IPC Technical Manual V3.1 Reference Table (IPC, 2021). Almost all (97%) of consensus classifications fall in the range of possible outcomes based on classifications implied by the underlying food security indicators. Notably, we present descriptive statistics that show that in the majority of cases, the underlying food security outcome indicators themselves are often discordant for a given area of analysis (or analysis unit) at a given point in time; that is, a range of potential consensus classifications are consistent with the Reference Table benchmark. This result suggests the important role of the consensus process in determining phase classifications and the percent of population in each phase.
Second, “What is the consensus process doing?” We find:

- TWGs use most of the available FSI data in determining the consensus outcome. That is, TWGs do not appear to disregard certain indicators (or elevate others), a concern raised by stakeholders. When we regress consensus phase and population estimate outcomes against the FSI derived population levels, we find that most FSIs exhibit a strong relationship with the final consensus outcomes.

- Our evidence suggests that the IPC analyses, on average, classify fewer people as being in IPC Phase 3 or above (crisis or worse) (i.e., is conservative) relative to the FSI reference points. Consensus outcomes are lower than the simple average FSI outcomes, and appear to more heavily rely on FSIs that imply lower outcomes. Further, we find evidence of a larger fraction of population in IPC Phase 3 or above (crisis or worse) identified as just below the cut-off of 20% than would be expected from a smooth distribution, suggesting that the consensus process is cautious with respect to classifying areas in IPC Phase 3 or above (crisis or worse).

Third, “What could be influencing consensus outcomes?” We find:

- The degree to which TWGs rely on particular FSI varies over time and space. Reliance on particular FSI also varies over the degree of severity of IPC Classification. Using regression techniques, we estimate the weights on each FSI for the classification and the estimation of the percent of population in IPC Phase 3 or above (crisis or worse) by (a) the set of available FSIs and (b) country and analysis round. We also compare findings to estimated weights (on each FSI for IPC Phase 4 or above outcomes. We show that weights vary substantially over all of these dimensions. The fact that the estimated weights vary is not in and of itself an indication of inconsistency of the process but does indicate that TWGs treat food security indicators differently across multiple dimensions, including over time in the same country. Thus, the same classification in the same country may be associated with very different levels of the FSI. This finding suggests that it may be misleading to compare consensus outcomes across TWG analyses within or across countries.

- We also find that the consensus outcome is more conservative when input data are more ‘noisy’ - i.e., when the underlying FSI data are either less likely to suggest the same outcome or when they suggest a wider range of outcomes.

Fourth, our AFI projection findings sought to answer: “What is the relationship between projections and current status classifications?” We find:

- For AFI projections, we compare projected phase classifications to the next period’s realized current phase classifications. We examine the degree to which projections tend to over or underestimate the phases compared to the realized current analysis for that period. We find that 4 in 10 forecasts are consistent with the resulting consensus outcome conducted in the next period. Under and overestimates are about equally split at around 3 in 10 each. Thus, while we find that projections often do not match realized current status analyses, we do not observe systematic over or underestimates of the projections. Circumstances are likely to change in unexpected ways, either due to new weather shocks, changes in conflict patterns or other unpredictable events transpiring in the intervening period. We also note that in crisis situations, one aim of the IPC projection is to generate a response. If projections do crowd in humanitarian funding, we might expect the realized current status to be lower than projections. Given under and overpredictions are equally represented, we do not observe a systematic decrease in phase classification from projection to realization.

Our AMN analyses were constrained by data limitations. We therefore present three example analyses using data from two country cases: Kenya and South Sudan. Given the small sample, we do not interpret our findings as indicative of broader results. Nonetheless, these analyses demonstrated three approaches to evaluating AMN consistency.
AMN current status and projection findings include:

- Specifically, we evaluate the reliability of evidence used for classifications and (1) between representative and non-representative survey-based outcomes; (2) between classifications implied by WHZ indicators and the actual consensus-based classification; and (3) between projections made for time t in t-1 and actual current status at time t.

In sum, our results underscore the vital role of the consensus process in IPC analyses. We make six main recommendations based on our findings to support fidelity to the IPC process and to improve comparability across classifications.

- First, given that the large majority of IPC analysis areas have FSIs that imply different phases at a given time, we suggest that the FSI thresholds in the reference table should undergo systematic study and review.

- Second, we recommend investigating whether automated analyses paired with documentation would support the consensus process, including TWG learning and comparability across and within countries. For example, it might be helpful to present TWGs with information about the correlations among the FSI data or a simple average or weighted average of the FSI-implied outcomes for each IPC analysis area to start discussions. While the consensus process may diverge from those initial algorithmic assessments, such starting reference points could provide TWGs with anchors for documenting when and why analyses deviate. This might help with transparency and consistency, and comparability of the process across TWGs. Note that this recommendation does not mean that we suggest replacing the consensus process with automated data analyses, only that additional analyses might complement the TWG process. Human-led consensus processes remain critically important, aiding both buy-in and coordination and reflecting existing limitations in algorithmic approaches and data availability.

- Third, we call for greater investments in data quality and consistency, not only within the IPC but within its partner organizations. For example, having access to the underlying SMART survey data would benefit both the consensus processes and future evaluations.

- Fourth, data quality reliability scores do not capture important aspects of data, such as possible sources of measurement error that could assist TWGs in their consensus process. Currently available data reliability scores reflect only timeliness and representativeness. We recommend investigating options to expand reliability scores to capture issues with data collection or other relevant information. Alternatively, documenting data concerns can help explain circumstances in which TWGs diverge markedly in their consensus classifications from those implied by the input data.

- Fifth, we recommend that the IPC partners continue to invest in and improve platforms to collect AMN data.

- Sixth, we propose that the IPC undertake a follow-on accuracy study in five years, when the availability of more longitudinal data could support additional in-depth analysis of the IPC’s accuracy.
1. INTRODUCTION

As the global standard for food security and nutrition analysis at national and sub-national levels, the Integrated Food Security Phase Classification (IPC) is being implemented in over 30 countries (IPC 2023). The IPC has been applied in 35 countries (IPC, 2019) and in 2021, the G7 called the IPC “the gold standard” for food security analysis (FCDO 2021; see also Buchanan-Smith et al. 2023). Rapidly becoming the dominant food security classification system globally, the IPC is used for early warning, determining the scope and location of humanitarian need and prioritizing resource allocation across crises. The IPC is also used for famine declaration, focusing global attention upon the most serious form of food insecurity. Because it assesses the scale and severity of food crises and provides classifications meant to be consistent and comparable across regions and time spans, IPC analyses guide the strategic allocation of six billion dollars of humanitarian food assistance per year around the world (IPC 2023).

The accuracy and integrity of IPC analyses are therefore of utmost importance for international and national humanitarian response. Understanding the specific conditions of time and place that can inform the accuracy of IPC classifications will support their interpretation and effective use. Even so, a system-level study to examine the accuracy of IPC classifications has not been conducted over the more than 15 years it has been in use.

Recognizing this need, the IPC global support unit in 2021 commissioned an Accuracy Study concept note which would set out basic parameters for such a study. This concept note envisioned a scope of work, which would include study of both acute food insecurity and acute malnutrition analyses for current status and projected classifications.

Determining how best to conduct such a study was a difficult problem, for a suitable, robust, off-the-shelf methodology was not self-evident. Food insecurity is itself a latent variable; it cannot be measured directly. Further, the IPC’s consensus-process based methods distinguish it from other food security analyses. Its classifications use a convergence of evidence approach based on aggregated, summarized household food security and or nutrition indicators and contributing factors to determine the degree of food insecurity experienced by analysis area-level populations (e.g., populations within districts). Drawing on these data and following a standard set of protocols, IPC TWGs determine both the classification of severity by analysis area and the population in each classification.

Ideally, one would compare IPC classifications (and the indicators underlying those classifications) to the actual experience of food insecure and malnourished populations -- that is, to the ‘truth’ of populations’ lived experience. However, there are no ‘true’ estimates of either severity or populations against which the accuracy of IPC outcomes can be assessed. Therefore, the extent to which IPC classifications and population estimates ‘accurately’ reflect lived experiences is unknowable.

Thus, the first phase of the research process was to develop a viable methodological approach to examining accuracy in the IPC context (see section 2.2.1). That stage was completed in April 2023, in a Scoping Study, and presented to the IPC TAG. During that stage, the research team completed key informant interviews to develop a clearer concept of what can meaningfully constitute accuracy for IPC as well as identified approaches to examining accuracy with the data available. The accuracy study represents the culmination of this research, the outcomes of this review process. We present and discuss findings with regard to overall accuracy of IPC analysis and we offer recommendations for improvements intended to support the work of IPC practitioners around the world.

1.1. Objective and approach: the role of consensus

Findings from our key informant interviews informed our approach to the question of accuracy. As discussed in greater detail in section 2.2.1, results from our interviews revealed that IPC stakeholders consider the IPC consensus process both a strength and a potential weakness.

The convergence approach at the core of the consensus process is fundamental to what the IPC describes as its “multi-sectoral cooperation and technical consensus”, a means of involving the interpretations of a range of stakeholders, incorporating context, overcoming data limitations in quality or availability, and achieving acceptance and ownership of the final classifications. Respondents consistently stressed the importance of the consensus process and cited its utility in incorporating contextual information.

Respondents also highlighted that the consensus process plays an important role in creating confidence in the outcomes for TWG members and their organizations.
However, respondents were concerned that the consensus process could also be a “black box,” to outsiders and could be susceptible to capture by vested interests or susceptible to implicit or explicit biases.

Respondents also raised concerns about whether TWG findings accurately reflect conditions across a national landscape, across a range of countries, or over a span of time. In other words, respondents were unsure if different TWGs would (or should, given the value of contextual knowledge) arrive at the same set of findings if they were given identical information.

Based on these interviews, methodological literature, and the data available, our analysis and our treatment of accuracy focuses on the consensus process. We examine IPC accuracy through the lens of fidelity to the IPC process and through the lens of comparability of outcomes across and within TWG analyses. That is, we examine whether TWG analyses are consistent with benchmark established by the IPC partnership at global level, as set out in the Technical Manual, and whether they are consistent with reference points drawn from the underlying food security indicator data. We draw on both the observed IPC analysis outcomes and the underlying data that go into the consensus process to reach our conclusions.

For acute food insecurity (AFI) analyses, we answer four broad questions, presented by increasing analytical complexity for the current status analyses, followed by our analysis of projections.

• First, we seek to understand what the underlying indicators suggest to TWGs about phase classification and population outcomes. We describe the nature of the data challenges including both the amount and quality of data available to a given TWG analysis and the degree of coherence across the indicators for a given analysis area at a given time. The lack of concordance among the FSI indicators points to the importance of the convergence process. In 40% of our sample, the range of classifications implied by food security indicators is two or more IPC phases. We characterize the degree of consistency in the use of food security indicators by TWGs in their population and classification analyses. We find that 97% of classifications are consistent with classifications implied by food security indicators. This suggests that TWGs play a substantial role in determining phase classifications and the percent of population deemed to be in each phase.

• Second, we analyze how the consensus process uses the food security information available. We compare consensus-based outcomes to outcomes estimated using an arithmetic (simple) mean of the FSI. TWG analysis outcomes are not simple averages. We find that IPC TWGs are likely under-classifying populations relative to the FSI reference points derived from the simple mean, on average. This finding is important in that it runs counter to the concern that we heard from the qualitative interviews that the consensus process might be pushed by some participants to generate higher phase classifications than otherwise warranted. On the other hand, if the process implies that some food insecure populations go unidentified, particularly in settings with bad data, that would be a serious concern.

• Third, we study what factors might influence consensus outcomes. We compare consensus-based outcomes to reference point outcomes using logit and OLS models. This analysis lets us examine whether certain food security indicators are systematically ignored or disregarded in TWG analyses, reflecting a concern raised during the Scoping Study phase that particular indicators had outsize influence. We do not observe this, on average. We do find that the treatment of FSIs in the consensus process is idiosyncratic and varies by severity of the phase classification, by country, by TWG analysis within country, and by the available subset of food security indicator data in front of the TWG. This finding raises questions about the comparability of analyses. When input data are noisy, the classifications tend to be more conservative, which suggests that the TWGs may be exercising caution in classifying analysis areas or populations in IPC Phase 3 or above (crisis or worse). In the future, TWG documentation of classifications that depart from modeled outcomes could build an evidence base about which certain indicators seem more or less useful in specific contexts.

• Fourth, we ask how AFI projections relate to later analysis. Specifically, we compare AFI projections to the actual classification for the projection period. We find about 40% of projections are consistent with the later current status analyses. We do not see evidence of projections systematically over or under predicting relative to realized current status analyses. Interpreting these findings is difficult. On the hand, a concern raised in the Scoping Study was that projections would be intentionally high to crowd-in resources, which we do not observe. On the other hand, if projections successfully induced a food assistance response adequate to generate interventions to promote a significant improvement, we would expect projections, on average, to be higher than later current status. This (lack of) finding raises questions for future research on whether and when projections induce an increase in aid and whether the increases keep populations from experiencing further deterioration or are enough to improve the overall situation.
Due to data limitations in acute malnutrition (AMN) classifications, we present a series of proof-of-concept analyses for AMN current status and projections. We note chronic food insecurity IPC classifications are beyond the scope of this work. We find severe data constraints underlie AMN analyses. Given that GAM analyses are required for famine declaration, improving the quality and availability of AMN data should be a priority.

As we show, IPC classifications and population estimates are not mechanical measures but are outcomes of decision process.

Understanding the accuracy of the IPC and modeling and analyzing the consensus process is a topical and important area of research not only for the IPC itself but also for policymakers and users of the IPC, including researchers. Because the IPC is nearly twenty years old and in a process of expansion to new countries, understanding the fundamentals of the analysis process and studying accuracy is timely and of particular importance. Moreover, IPC analyses are increasingly being used in research as sub-national measures of food security. A paucity of data on sub-national food security that is spatially and temporally comparable make the IPC analyses particularly attractive to researchers looking for food security measures to use as dependent and independent variables in analyses linking food security and changes in food security to climate change, conflict, and national and international policies.

1.2. Roadmap

The rest of the report continues as follows: In Section 2, we discuss our methods for evaluating the current and projected AFI. This section discusses the methodological challenges and then presents a model of the IPC decision-making process and various statistical approaches to explore that process and its relationship to the underlying input data to explore that model. In Section 3, we present our data, including summary statistics of the outcomes and inputs. In Section 4, we present the results of our analyses on AFI data. In Section 5, we discuss the methods, data, and results of our AMN analysis. Section 6 highlights our findings and study limitations, and Section 7 ends with recommendations. We present further analyses in the Annex.
2. NATURE OF THE PROBLEM AND METHODS

2.1. IPC and food security: Nature of the problem of analyzing accuracy

What, in the context of the IPC, and in the broader context of food security measurement and analysis, is meant by ‘accuracy’? Ideally, we would compare an IPC classification of an IPC analysis area and time period to the lived experience of food insecurity and malnutrition experienced by all people living in that location at that time. However, unlike other concepts, no single, observable, regularly collected metric captures individual food security or malnutrition. Food security is by definition a multidimensional construct, covering aspects including the quality and quantity of nutritional intake and related to household behaviors undertaken in response to a lack of accessible food and nutrients. Similarly, an individual can present as malnourished in multiple ways. IPC’s convergence of evidence process can incorporate the multiple dimensions of these latent concepts, while also allowing for contextualization.

The latency of food security and malnutrition pose a fundamental challenge in that there are no observable, objective measure of ‘truth’ against which we can compare the IPC malnutrition and food insecurity classifications. Thus, we cannot ‘validate’ the IPC per se. Instead, we explore whether it systematically varies with other measures of food security and malnutrition over time and space. The IPC itself is a form of food security index; TWGs aggregate available food security measures (as well as nutrition measures when multiple forms of data are available) and contributing evidence using a convergence of evidence approach. In other words, TWGs implicitly value some indicators and some evidence more than others in achieving consensus. This valuing is akin to TWGs implicitly applying different weights to pieces through the TWG consensus process. We clarify that TWGs are not explicitly weighing, rather we observe outcomes of consensus that use a convergence of evidence approach. Statistically, we can estimate the implicit weights.

2.2. Evaluating metrics when there is no objective, measured ‘truth’

Social science researchers and policy makers are often concerned with the measurement and tracking of latent constructs central to human welfare including poverty, hunger, agency, and empowerment (Sen 1983, Barrett 2010, Bourguignon 2006). These concepts share a common feature: they are not directly observable and measurable. This distinguishes them from readily measurable indicators including height, weight, income, prices, and disease incidence. Measures of latent constructs are therefore built from other, more readily assessed variables.

Given that measures of latent constructs are imperfect proxies for the true unobserved concepts, an active literature focuses on measurement: how to (for example) enhance what the measures of poverty and food security encompass, and how to add dimensions or incorporate subjective and objective analyses into their construction and interpretation (Alkire and Foster 2011, Ravallion, 2016, Charmes and Wieringa, 2003, Duflo, 2012, Phradan and Ravallion, 2000). A second literature proposes new measures for latent constructs; researchers have developed a range of new metrics to assess concepts including water security (HWISE, see Young et al. 2019), women’s empowerment (Alkire et al. 2013; Narayanan et al. 2019), and food security (Clapp et al., 2022, Herforth et al. 2020).

It is not possible to assess the accuracy of measures of latent constructs, and so scholarship focuses on validating new and existing measures generally explores the consistency of measures with one another or whether they capture the same populations, and in some cases, whether they capture the same latent concept (Maxwell et al. 2014, Klasen 2000, Headey and Ecker, 2017, Gebreyesus et al. 2015). For example, Vaitla et al. (2017) use rich household-level surveys including combinations of four measures of food insecurity (FCS, HDDS, HH, and rCSI). They analyze relationships between measures (e.g., bivariate correlations) and apply factor analysis in an effort to identify latent (unobservable) dimensions of food insecurity. They find that food insecurity dimensions include both a lack of quantity of food consumed (well captured by HHS and rCSI) and a lack of diversity of food consumed (well captured by FCS and HDDS); other aspects of food insecurity may exist beyond these two. They propose that best practice food insecurity data collection should include both a quantity (HHS or HH) indicator and diversity (FCS and HDDS) indicator.

In contrast, one recent study (Maxwell et al. 2023) attempts to ground-truth classifications using the Household Hunger Scale (HHS) to differentiate between households in Phase 4 (Emergency) and Phase 5 (Famine) using extensive interviews and observation by enumerators. This is one of the few (if only) studies that starts with IPC classifications and seeks to identify characteristics of households within those classifications.
The IPC is both a measure but also, in its assigning of a phase and population figures to an analysis area, a metric; it is used by governments and international agencies for targeting and intervention. It is forward-looking and intended to crowd-in aid. A second relevant challenge with respect to accuracy therefore relates to the fact that the forecasted food insecurity classifications themselves change future states by influencing aid allocations and, in some cases, even reducing the severity of need in the next period. As one respondent in the qualitative round of our research pointed out, if the projections are effective in terms of marshaling and distributing aid to places that need it in time, they will never be correct. In this way, evaluating the accuracy of IPC projections presents challenges similar to evaluating the accuracy of other early warning programs including those working to reduce deforestation (Finer et al. 2014), or to mitigate the spread of disease (WHO 2011).

2.2.1 Qualitative findings

In order to understand accuracy for the IPC, we gathered reflections regarding what accuracy means to the IPC community. We then used these observations to inform our working definition of accuracy and to identify key quantitative research questions. Between October 2022-January 2023, the research team conducted interviews to identify and understand perceptions of (in)accuracy in the IPC. Based on a sample drawn from IPC GSU global staff, regional coordinators, FRC, TWG, and TAG members, as well as other users in IPC partner and donor organizations, the team conducted 21 interviews with people highly knowledgeable about IPC.

The intent of this qualitative research phase was to cast a wide net over a broad range of possible concerns to inform our approach to the accuracy study. This approach was intended to create composite sketch of the overall range of concerns, without pre-emptively delimiting the scope of quantitative research. However, it was not intended to generate a stand-alone qualitative analysis of IPC accuracy, in part because it is not feasible to differentiate between perceptions of inaccuracy and accuracy itself. Interested readers may wish to read the Scoping Study.

Ultimately, the interviews helped:
1. to inform the study team more fully understand what (in)accuracy might mean for IPC
2. to identify where inaccuracy enters into (or is believed to enter into) the IPC process and
3. to identify possible drivers of (in)accuracy that could be studied using quantitative tools.

The emphasis of these interviews was on gathering and collating perceptions, highlighting heterogeneity where it was identified.

Qualitative interviews identified several moments when inaccuracy could enter the IPC process. We sorted the findings on the nature and driver(s) of (in)accuracy by the IPC’s four functions (1. Build Technical Consensus; 2. Classify Severity and Identify key drivers; 3. Communicate for Action; and 4. Quality Assurance). We focus here on function 2, which is where the majority of concerns related to inaccuracy were centered, and where quantitative tools could be applied. Specifically, much of the focus of the qualitative interviews was on the TWG-level consensus process and whether consensus outcomes were consistent with IPC guidance. We also note that some issues (e.g., a concern that limited or weak training of enumerators on data collection) falls outside of the IPC’s purview, and therefore were not considered within the scope of this research.

Several respondents described the IPC as a gold standard for its role in food security and malnutrition analysis. Respondents acknowledged that the IPC is highly valued, with respondents stating “We take it for granted. Other sectors are jealous” and “Without IPC data, the [humanitarian] world stops”. Stakeholders reported relying on the IPC for early warning, to identify need, as a scorecard for understanding whether funding is making a difference, and for prioritizing across and between crises. TWG analysts were seen as deft and well-informed.

Concerns were also raised, with one respondent noting “The system is quite amazing actually, given all the vested interests”- vested interests, in this context, referring to parties with a stake in what classifications emerge from an analysis. Elsewhere, the IPC was described as “[a] victim of its own success”: by virtue of its widespread adoption, the IPC is used in ways not intended (e.g., as a scorecard for whether interventions impact classifications) and is expected to incorporate and report on more and more information.

---

1 Research has shown that robust evidence is the strongest protection against the politicization of humanitarian need and humanitarian analysis (Maxwell and Hailey 2021). Robust evidence is also critical to addressing misinformation more broadly and to improving judgements (Kahneman et al. 2021).
Below is a selection of issues that informed our approach to studying the accuracy of the IPC. Interested readers may wish to review the Scoping Study, which served as the basis for the main quantitative findings, which are presented in Chapters 3 and 4.

On consensus

The consensus process which underpins the IPC was viewed as both a strength and a potential weakness. It was seen as useful for incorporating contextual information and encouraging buy-in from stakeholders, but it was also described as at risk of potential bias and, in some cases, capture, by vested interests (see above). Some perceived that results could be influenced by individual participants’ roles and reputation, race, gender, institutional affiliation, etc. The ‘loudest voice in the room’ could come to dominate discussion (for more on this see, ATARI; Buchanan Smith et al. 2022). TWG chairs or strong members may take advantage of their position to benefit from a particular outcome and seek to maneuver the analysis to attain that result.

The politicization of findings was commonly mentioned as a potential driver of inaccuracy, with examples drawn from a few complex crises raised repeatedly. In these analysis areas, drivers of inaccuracies were seen to be more prevalent in analysis of active conflict zones, especially when the government is a party to conflict. Given that the same few examples were raised numerous times, it may be that concerns about the consensus process itself as a potential source of inaccuracy reflect a small number of cases which weigh heavily in respondents’ experience.

For those not in the room at the time of the analysis, consensus processes could appear as a ‘black box’ for some end users, who may be unclear on how, or why, a particular classification was made.

On convergence of evidence

Convergence of evidence refers to the process through which data are brought together in the IPC analysis process, accounting for data reliability, consistency, timing, and many other factors. While convergence provides a basis for reconciling differing opinions, respondents were uncertain if classification thresholds included in the reference table were ‘correct’ - that is, if the classifications thresholds were set at the appropriate level. For example, some respondents were uncertain as to whether the cutoff thresholds for individual indicators were set too high or too low in the Reference Table or were too high or too low for their specific context. In application, there were questions if these thresholds were then applied consistently across TWGs.

Opinions among interviewees varied regarding which indicators were better than others, but there were no stand-out ‘best’ or ‘worst’ indicators highlighted: all IPC outcome indicators had supporters and detractors. Concern was raised that some of the metrics were applied inconsistently across locations. Further, some indicators point to very different classifications, and respondents were concerned that particular indicators would be dismissed or would be used in contextually inappropriate ways. A few respondents also speculated that given that indicators can point to different classifications on how accurate an ‘averaged’ outcome classification would be. This concern reflects an ongoing conversation in the humanitarian space about the roles of qualitative analysis, expert opinions, and machine-learning or algorithmic approaches (Lentz and Maxwell 2022). These concerns suggest there could be variable and or inconsistent weighting of indicators across and within TWGs.

At the same time, it may not be reasonable to expect consistent indicator weighting across TWGs. Several respondents argued that some food security first-level outcomes (including livelihoods coping strategies and rCSI) may be more applicable in some contexts than in others (e.g., one respondent explained that eating wild foods is an indication of adverse coping strategies in some regions, while in others it is not. The livelihood coping strategies index asks whether wild foods “normally not eaten” are being consumed but whether this nuance was effectively conveyed in household surveys was unclear to the respondent). An implication of this is that similar food security first-level outcomes in two contextually different analysis areas may result in somewhat different classifications. It was unclear to some respondents whether, how, and in what cases, TWGs account for context when evaluating first-level outcomes, which feeds into perceptions that consensus is a ‘black box’ (see also Famine Review Committee 2022).
On population estimates

Qualitative interviews suggested that there were more concerns about the accuracy of population numbers than the accuracy of phase classifications. Some respondents suggested that the classification (i.e., Phase 3) was correct but the population estimates were too high. For example, 40% of the population might be classified in Phase 3; respondents suggested that while the area was likely in Phase 3, maybe only 20-30% of the population was actually experiencing Phase 3). One possible reason given for these higher numbers is a suspicion that higher populations estimates crowd in resources and or attention. Other respondents suggested that the population estimates from the previous analysis exercises often implicitly anchored the current discussion (i.e., are more or less people food insecure than were in the previous analysis exercise?). In other words, prior population figures act as starting points which delimit the ‘acceptable’ range of increases or decreases possible to be realized across the five-point phase in subsequent analyses.

Emerging questions

Some of the questions emerging from these interviews included:

• Are TWGs using all available information? Is there a tendency to rely on (or ignore) only one food security indicator or to simply average them?

• Are TWGs over-estimating populations in IPC Phase 3 or above (crisis or worse)?

• Are TWGs projections overly dire (i.e., to crowd-in funding)?

• Are consensus outcomes comparable across countries?

These questions helped to refine the scope of research. We answer these questions for AFI and show proof-of-concept approaches for AMN.

2.2.2 Defining study

Findings from the qualitative interviews, our review of the literature on examining the accuracy of latent constructs, and our evaluation of the data available at GSU helped prioritize our analyses. Findings from the Scoping Study helped us to arrive at a definition of internal consistency. Observations on the latency of the IPC and indeed of underlying food security indicators resulted in clarifying that accuracy, for this study, differs from validity: validation refers to the extent to which the object of study measures what it is intended to measure; and in the case of food security, this cannot be observed.

To examine the accuracy of AFI and AMN current status and projections, we therefore focus on whether consensus outcomes are consistent with our benchmark, the IPC Technical Manual V3.1 Reference Table (2021) guidance (see Sections 2.3, 3, and 5 below). This is a measure of internal consistency. When possible, we also include additional analyses that do not rely on assumptions about the reference table; drawing on statistical properties, these analyses act as reference points (see Sections 2.4 and 4.1 below). For example, to understand the underlying distribution of TWG population outcomes, we compare them against a reference smoothed distribution. This reference distribution shows an estimated population distribution that ignores population threshold effects.

To study internal consistency, our approach focuses on evaluating the IPC decision-making process and comparing the outcomes of the process to the input data used in their development. We break the problem into four parts, presented by increasing analytical complexity for the current status analyses, followed by our analysis of projections.

• First, we ask “What do the underlying indicators suggest to TWGs about phase classification and population outcomes?” This question seeks to explore the information TWGs have.

• Second, we ask “What is the consensus process doing?” This allows us to examine if population or phase classification outcomes are systematically biased.

• Third, we ask “What could be influencing consensus outcomes?” This focuses on what could be shaping the outcomes we observe.

• Fourth, we ask “What is the relationship between projections and current status classifications?” This allows us to examine if projections are systematically higher or lower than consensus classifications.

These questions provide a way of addressing many of the concerns raised during the Scoping Study about how consensus happens and whether particular biases (e.g., to overestimate populations in need or to generate alarmist projections) exist.
2.2.3 Study parameters

Our study design was informed by several parameters.

First, we do not have access to raw data nor do we have access to notes or conversations that occurred during the TWG consensus process. We do not review contributing factors, which are qualitative in nature, nor do we incorporate contributing factors into our analyses. This could be a valuable area for future work as these information sources become more readily usable in quantitative work (i.e., through machine readable formatting, etc.).

Second, we for purpose of this study, we define as benchmark the IPC Reference Table (2021). In our recommendations, we propose that there may be merit in examining food security indicator thresholds given the level of discordance we observe across food security indicators but for the following analyses, we take it as given. We compute several reference points based on the values in the IPC Reference Table (e.g., taking the simple mean of phase classifications using food security indicators).

Third, the majority of AFI analyses draw from a sample of 1881 TWG classifications across 15 countries during the end of 2020 through 2022 that include food security indicator information or a subsample of these classifications. These countries include the set with data housed at the GSU that were machine readable and shareable. The sample may not be representative of the wider set of IPC countries. When underlying food security indicator information is not needed in the analysis, we use larger samples of machine readable TWG classifications and population estimates from published IPC reports during the end of 2020 through 2022. We round population estimates to the nearest 5% for consistency with TWGs’ approaches. Across both types of samples, we have very few cases of classifications of phase 5 and therefore our findings should not be extrapolated to phase 5 contexts. We indicate which data sets are used in the captions under each figure and table.

As a result of both the nature of the problem and the nature of the data at hand, when our findings differ from consensus outcomes, we cannot conclusively isolate the driver for the difference nor can we determine which measure is more or less accurate. The cause of inconsistency could reflect limitations in our reference points or models. Alternatively, the inconsistency could reflect that consensus is working; for example, if the TWGs adjust outcomes due to data we cannot observe. Alternatively, consensus may not be working, for example the loudest voice in the room may exert outsize influence. We return to documentation in the recommendation section, proposing that documenting major deviations from the Reference Table or other reference points could be valuable for understanding.

2.2.4 Note on language

As described above, we cannot definitively identify the “accuracy” of a latent concept, such as the IPC. We underscore that the language we use is intended to clarify rather than impose judgement on the findings. For example, when we use the term “conservative” we intend this to mean a finding is conservative relative to the reference point. We acknowledge that a reference point itself may not be correct and an alternative and equally plausible interpretation is that the reference point is incorrect. Nonetheless, we operate from a definition of “internal consistency” for this study, we taking as given the reference points we use. As we discuss in the recommendations section, future work could interrogate whether the reference points are themselves accurate.

2.3. Conceptual framework to understanding consensus

To evaluate the consensus process, we first need to understand how it is intended to work. The IPC consensus process involves a series of steps to reach agreement among analysts within the TWG (IPC, 2021). It begins with the guidance of a qualified facilitator who helps define ground rules for building consensus. Analysts walk through the IPC analysis worksheet to analyze and discuss evidence related to their sector(s) and the IPC protocols, aiming to achieve a strong understanding of the underlying data, and to reach agreement around both the classification of IPC analysis areas and the proportion of population in each phase classification.

We appreciate the suggestions by Tim Hoffine and Carlo Cafiero on this point.
The IPC Analysis Worksheet is organized into 12 steps. Overall, reaching consensus within the TWG involves critical analysis, discussion of evidence, addressing disagreements through neutral facilitation, seeking agreement at the country level, documenting minority views when necessary, and considering external quality reviews for classifications 4 and 5 (IPC, 2021). An external quality review may be requested by members of the TWG or supporting partner(s) reflecting the minority view (IPC, 2021). Vetting of classification and population estimations is also recommended as good practice for IPC consensus-building (IPC, 2021).

For acute food insecurity classifications, steps 1 and 2 of the AFI Worksheet involve identifying the context and analysis parameters and focus on populating the evidence repository that is created to organize a wide range of evidence from multiple sources for all classifications (IPC, 2021). In step 1, the TWG examines the area’s characteristics and the households within it, allowing for the contextualization of evidence and livelihood-based analyses. In step 2, the TWG reviews a comprehensive range of data details, which are collected in the evidence repository. This evidence encompasses references for all the evidence earmarked for analysis, encompassing source details, collection dates, and reliability scores, thus facilitating easy access and reference to the evidence throughout the analysis process.

In Step 3, the TWG examines the evidence for the current classification using the IPC Acute Food Insecurity Analytical Framework and Reference Table from the IPC V3.1 Technical Manual (2021). This examination involves considering the local context, evidence reliability scores, historical trends, and socio-economic differences. The consensus process uses various food security indicators, including the food consumption score (FCS), reduced coping strategy index (rCSI), livelihood coping strategies (LCS), household hunger scale (HHS), and food insecurity experience scale (FIES), while factors such as food production, food source, income source, weather information, and humanitarian aid programs are included as contributing factors.

In Step 4, the TWG focuses on determining the phase classification and estimating the population for the current classification. The TWG classifies units of analysis based on the severity stage experienced by the worst (at least) 20th percentile of the population. This means that even if the overall population distribution differs, the focus is on identifying the severity level that affects the most vulnerable segment. For example, if a population is distributed as 50%, 20%, 10%, 10%, and 10% across increasing severity phases from Phase IPC 1 to IPC 5, applying the classification threshold of 20% would classify them as IPC 4, since 20% are categorized as IPC 4 or 5. Similarly, if the distribution is 5%, 5%, 5%, 80%, and 5%, the same classification of IPC 4 would be applied based on the rationale of prioritizing the worst 20th percentile.

In Steps 5 and 6, the TWG aims to identify areas that have received significant humanitarian food assistance and comprehend the main drivers and factors that restrict food security within the current classification. In our analysis due to data limitations, we rely on the key outcome indicators (food security indicators) to examine the convergence process. All these pieces of evidence are recorded in the Consolidated Data Tool (CDT); additional information on this can be found in Annex 1 for more specific details.

The remaining steps, 7 to 12, continue the analysis for projection classifications, including developing assumptions for future shocks, analyzing evidence, identifying areas of planned humanitarian food assistance, and determining priority strategic response objectives.

Consensus is not always achieved, and disagreements may arise in specific areas or in the overall analysis. In such cases, the analysis team uses neutral facilitation to address disagreements and to seek agreement at the country level to minimize delay in releasing time sensitive findings (IPC, 2021). If agreement cannot be reached, dissenting organizations can choose to disagree with the analysis results, with a range of options existing for expressing less than full agreement with the final results, although cases are rare (IPC, 2021).

The AMN process is similar to AFI with a few exceptions. While both analyses involve multiple steps, AMN focuses on two distinct approaches based on the type of evidence analyzed: nutritional outcomes (specifically WHZ/MUAC) and other contributing factors in Step 3. The identification of areas receiving humanitarian food assistance is excluded in both current and projection analysis for AMN. Furthermore, the AMN worksheet does not include “identification of risk factors to monitor”, which is included as Step 11 in the AFI worksheet.
2.3.1 A simple model of the consensus process

For both categorization and determination of population shares, we model consensus of any TWG in any round as a process of weighing the input data, and then making adjustments to the resulting IPC level and population shares to conform to TWG members’ intrinsic analyses of the local situation.

\[ y_{it} = \beta_t \text{FSI}_{it} + \mu_i + \alpha_{it} + \epsilon_{it} \]

where \( y_{it} \) is the IPC analysis for IPC analysis area, \( i \), in round, \( t \). The vector \( \beta_t \) are the weights developed by TWG on the food security indicators (FSIs) for the entire country at time \( t \). These weights may come from the fact that TWG members are skeptical about the relevance of a food security indicator for their country (e.g., LCS), or how specific survey questions were asked (e.g., self-reported questions about perceptions of hunger). If the concerns are around the veracity of a specific measure for the country, we would expect the process to result in similar weights on that FSI across the country. TWGs may also adjust the categorization for a specific location for regions that may be particularly vulnerable or have characteristics that make the population weighted FSI measures less relevant (e.g., refugee camps), represented by \( \mu_i \). If the characteristics of these analysis areas do not change (greatly) over time, we would expect this adjustment to be similar across rounds. TWGs may make adjustments for a specific IPC analysis area in a specific round, \( \alpha_{it} \) due to local information that is not captured in the FSI input data. Last, \( \epsilon_{it} \) are random errors. As noted above, we assume a similar model for both classification and for the determination of the percent of food insecure population in each Phase.

2.4 Application of methods for current status AFI

Our qualitative research identified a major source of concern: consensus-led processes may result in IPC classifications that differ from ‘true’ results, especially for population estimates. While ‘true’ results cannot be observed, we compare the underlying food security outcome data to the observed classification data that might allow us to identify internal consistency.

Comparing the IPC across analysis areas and time is challenging for several reasons. The subset of indicators varies by country, and the quality of the underlying data will also vary by location and time. Some indicators may be better at identifying food insecurity for different populations or analysis areas. Relatedly, some indicators are better at capturing different severities of food insecurity. Combined, this means a priori that certain variables are more helpful in certain contexts. So, we might not expect to see internal consistency across TWG analyses. But we might expect TWG analyses for a specific country to be consistent in their treatment of indicators, although we recognize this may be affected by data quality and context. Further, we recognize that some indicators may be better at capturing different levels of severity than others; thus, we might expect that if two IPC analysis areas have different levels of severity, they may also have different underlying weights on the FSIs. Further, it may be that unobserved shocks are correlated with some but not all food security indicators, and that this may vary by context.

2.4.1 Arithmetic (simple) mean

Results in §4.1.2

We examine differences between a) the observed IPC 3+ consensus population and b) the population averaged across food security indicators that meet the IPC 3+ classification requirements. The observed IPC 3+ consensus population is the observed population requiring urgent action, which is the sum of reported populations in IPC 3, IPC 4, and IPC 5. This value is derived from the population tables in the IPC reports.

The population averaged across food security indicators acts as a reference point, which allows us to understand the size of the difference between the observed population classification derived through consensus and the averaged population values. In other words, we can identify how important the consensus process was in arriving at the observed population estimates. In the findings that follow, we define the impact of consensus as the area between the average of the food security indicators and the observed IPC outcomes.
We choose this approach because the arithmetic (simple) mean is straightforward to compute and relatively easy to understand. Further, this approach provides an internal consistency check by allowing us to understand the impact of the consensus process across TWGs, across combinations of different indicators, and across reliability scores. However, this approach has some limitations. First, the Reference Table is our benchmark, and we assume it is correct. Second, we caution that our reference point of taking the average across indicators is not following consensus protocols (IPC 2021) and could itself result in inaccurate results.

The AFI Reference Table maps a range of food security values for each FSI to an IPC phase. For each food security indicator, we use the classification table to classify the percent of population in each phase. We then sum the percentages of the population that meet the IPC3+ classification for each food security first level outcome indicator (FSI), i.e., rCSI 3+, HHS 3+, FCS 3+, and LCS 3+. For example, rCSI 3+ is the percent of population with an rCSI score of at least 19, which includes IPC 3, IPC 4, and IPC 5. We then take the arithmetic mean (the simple average) of the population requiring urgent action based on all available FSI 3+ population estimates.

While this approach is contrary to IPC guidance against using simple averages of FSI, we use it here as a point of reference and to identify the role consensus plays in population classifications. The simple mean should not be interpreted as more or less accurate than TWG-led analysis classifications of 3+, but it does support the identification of patterns, and allows us to explore how TWGs weigh different indicators in practice. Wide variation may also give rise to concerns about internal consistency. For example, if a TWG adjusts many population estimates away from the levels suggested by FSIs, that may be a result of important insights into local context, it may suggest that the working group is merely reporting its priors, or there may be other reasons.

### 2.4.2 Estimating weights

**Results in §4.1.3 - §4.1.5**

Next, we use this model to estimate the weights a TWG places on different FSIs. Note that we do not suggest that a TWG explicitly allocates weights to each FSI. Rather, the process results in implicit weights. We argue that these weights might be informative about the inner workings and consistency of the convergence process. We might expect that some weights are common to a specific TWG analysis and round, while others may be common to a location over time. For example, some TWG participants might bring to the analysis unobservable information that may be correlated with a specific FSI, implying that a specific FSI should receive higher or lower weight. Similarly, we may expect some TWG participants have knowledge of or interest in specific analysis areas, and thus argue that phase level be adjusted upwards or downwards compared to what would be predicted by TWG-level regression models.

We use this approach to understand whether the weights vary across TWG analyses, which might suggest that TWGs are using FSI information differently. Differences in weights may reflect a well-functioning convergence process. For example, TWGs may adjust their implicit weights to reflect that data quality or relevance varies over time and space. However, differences may also give rise to concerns about inconsistency across space and time across the set of analyses. We also use this process to identify whether the IPC generally under or over predicts the level and number of people who are food insecure relative to other food security metrics.

We conduct this analysis for both phase classifications, using two sets of logistic regressions, comparing the categorizations of phase 3 and higher, and phase 4 and higher against the percent of population in IPC Phase 3 or above (crisis or worse) classified by each available FSI. We run these logistic regressions by groups of TWG analyses using the same FSI inputs. We use the logistic regressions to compare the weights on the different combinations of available FSIs and by severity level. Since different FSIs capture different dimensions of food insecurity, we might expect that the weight on each might vary when they are grouped differently (e.g., we might expect a different weight on rCSI when LCS is included versus when it is not). Second, given that some FSIs are more or less relevant for different phases, we might expect that their implicit weights may vary by severity of food crises and thus by phase classification.

Because phase classifications can mask large variations in the breadth of food insecurity in a location (Andree et al (2020); discussed below in §4.1.6), we also regress the percent of population classified as in IPC Phase 3 or above (crisis or worse) (considered “in urgent need”) using an ordinary least squares regression against the population in IPC Phase 3 or above (crisis or worse) identified by the different FSIs. We use these regressions to explore how the weights on the different FSIs vary across rounds within countries, and to generate predicted levels of the population in IPC Phase 3 or above (crisis or worse) to compare against consensus outcomes in the following analyses.
2.4.3 Distributional analysis: “bunching” 

Results in §4.1.6

We explore the evidence of relative over versus underprediction associated with certain levels of severity. We compare the distribution of the percent food insecure to a smooth distribution to ask whether we observe “bunching”, i.e., larger than expected concentrations of observations at the cut-points between IPC categories. If the percent of population that is food insecure is randomly distributed, we would not expect to see mass points at specific levels like 19 or 20%. We might expect ‘bunching’ just below classifications that induce aid if, for example, a Working Group has insufficient data to feel comfortable calling for a response, even if other data might suggest that response is warranted. Further, one might anticipate that in some settings, even with sufficient data, political pressure moves IPC analyses to one side of a category or the other – either to induce aid or to avoid the scrutiny associated with aid that, say, a phase 5 or phase 3 might bring.

Specifically, we consider the distribution of population in phase 3+ and 4+ at the threshold between phase 2 and 3, and between phase 3 and 4 for the complete sample. We also use a predictive model based on weights on FSI estimated by TWGs, comparing the distribution of phase classifications suggested by this model with the distribution of phase classifications generated by the consensus process for individual countries. We examine whether there are differences in these distributions and whether there is evidence that the underlying population numbers cluster at just above or below IPC classification threshold levels.

2.4.4 Residual analysis

Results in §4.1.7

For three countries where we observe multiple rounds of analyses, we estimate the weights for each country and round for the FSIs using a linear regression and extract the residuals for analysis. These residuals can encompass all unobservable adjustments, whether or not they vary by place and time, and whether or not they are explicit in the TWG analysis process or mere error (noise): \( \mu_i + \alpha_{it} + \varepsilon_{it} \). Thus, these residuals may include specific location effects, and other adjustments by location and time. We explore whether these residuals suggest systematic adjustments from the weighted FSI mean. The weighted FSI mean can be thought of as a result of a TWG analysis process which places common weights on FSIs across the entire country.

For example, as above, if TWGs believe the livelihood coping strategies indicator is not relevant for a specific country, the TWG may implicitly weigh it less in its decisions. The deviations from this weighted average can be thought of as specific adjustments made to population numbers in a specific region. Through this analysis, we can observe whether residuals have systematic patterns over space and time or appear random, which might be consistent with them merely being noise.

2.4.5 Noisy input data

Results in §4.1.8

We use the above regressions to explore how the residuals vary when we have more or less divergence in the phase classifications suggested by the FSIs. We would expect that the consensus process would have a greater impact on the suggested phase classifications when the underlying data are more divergent, or ‘noisy’.

For example, if all FSIs suggest that a location is in phase 3 one would expect that both the consensus process and the regression generated weighted mean of the FSIs would generate a classification of phase 3 for that location. On the other hand, if half the FSIs suggest phase 2 and the other half suggest phase 4, while the regression-based outcome might suggest phase 3, one would not be surprised if the consensus process would result in anything ranging from a 2 to a 4. We consider both the number of FSIs that are in agreement along with the range of disagreement, both in terms of the percent of population in 3+ as well as in the classifications, as measures of ‘noise’.
2.5 Methods for projected AFI

Results in §4.2

One approach to examining projections is to ask whether the projections match the actual outcomes by comparing current status analyses to the projected status for that same period. However, there are limitations to this approach. A lack of a consistent match does not mean a poor projection was made for a few reasons.

First, food security analyses like the IPC can - in fact are intended to - change the future. We would hope donors, states, and others respond to projections, leading to lower current status analyses in the next period. To examine this would require detailed humanitarian food assistance (HFA) information, including on timing of delivery, which we did not have access to at the time of analysis.

Second, change happens for unexpected reasons; TWGs cannot see the future. That is, while they can make assumptions about the direction of various factors influencing food insecurity, they cannot anticipate all events (e.g., Covid-19; the Russia-Ukraine war) that have global or local impacts on food insecurity. Therefore, we warn that any documented discordance does not necessarily mean the projection analysis was wrong; rather, we are looking for average patterns, such as whether the projections are consistently above or below the realized current status. Finally, there are limitations related to timing. We can only compare projections that overlap with current status for that projection window. In practice, this meant that some projections could not be considered.

We also compare the distribution of the projections and the current analyses to ask whether one is more likely to exhibit signs of bunching. As noted above, bunching in either measure might suggest a tendency to either highlight or downplay potential current or future crises. One could envision tendencies in both directions and timing: one might want to stress potential future need to induce resources but downplay current analyses to avoid scrutiny.

We discuss methods for acute malnutrition below in §5.1.
3. AFI DATA AND SAMPLES

3.1. Data collection

We retrieved acute food insecurity (AFI) outcomes from the IPC-API. This includes information such as the total population analyzed, the number and percent of people falling into each phase outcome (ranging from 1 to 5) by country, analysis period, and area. These figures are available for the current period, as well as the first and second projections (projecting one and two quarters ahead) when produced. The data includes the IPC phase value (ranging from 1 to 5, as specified by the IPC Technical Manual V3.1) based on the population share. We further obtained geometric coordinates from the API to visually present our analysis results on a map in §4. For our following analyses, an observation is the classification of an IPC region in one round.

Our use of AFI outcomes involves two distinct approaches. First, we assess the convergence process by using a subset of the sample that has corresponding information on underlying food security indicators (FSI) obtained separately from the Consolidated Data Tools (CDT). Second, for analyses that primarily rely on IPC phase outcomes and population estimates, such as identifying bunching (§4.1.2) and comparing projection (t-1) to realized current status outcomes (§4.2), we extract specific cases from the entire set of classifications accessible through the IPC-API. Detailed information on the sampling procedures and outcomes for both approaches can be found in Annex 2.

3.2 Descriptive statistics – current status AFI

3.2.1. Sample size, FSI availability and FSI concordance

We begin by summarizing the food security indicators used in TWG classifications (TWGs classifications vary by country and by round). Table 1 provides an overview of the primary sample used in our analysis, including the sample size and availability of FSI. Our sample consists of data from 15 countries, with a total of 28 TWG analyses and 1881 classifications. As discussed in Annex 1, the initial sample consisted of 1901 observations; we excluded 20 cases that fell outside the 1 to 5 range of phase classifications; these 20 cases were either not analyzed by the TWG or deemed by the TWG to have inefficient evidence. We note that the proportion of total cases available that were not usable is quite small. FCS, rCSI, and LCS are most commonly collected. In our sample, we had few cases of FCS with lower cut-off (L_FCS) and FIES being used; therefore, we exclude them from our analyses. We provide the full description of sample size and FSI availability by analysis in Annex 2.

Table 1. Summary of cases and FSI availability

<table>
<thead>
<tr>
<th></th>
<th>L_FCS</th>
<th>FCS</th>
<th>rCSI</th>
<th>HHS</th>
<th>LCS</th>
<th>HDDS</th>
<th>FIES</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td># TWG Analyses</td>
<td>3</td>
<td>28</td>
<td>28</td>
<td>17</td>
<td>27</td>
<td>16</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td># of Classifications</td>
<td>135</td>
<td>1855</td>
<td>1844</td>
<td>1104</td>
<td>1823</td>
<td>930</td>
<td>56</td>
<td>1881</td>
</tr>
</tbody>
</table>

This table summarizes our sample of TWG analyses and classifications, including the availability of food security indicators (FSI). The sample consists of data from 28 TWG analyses with a total of 1881 classifications that have corresponding information on underlying FSI.
Figure 1 shows that 64% of the classifications are IPC 3, by far the most common AFI classification. As described above, our AFI sample is limited to recent data available to the GSU. The sample may include more populations in IPC Phase 3 or above (crisis or worse) than the universe of IPC classifications. Our sample includes no phase 5 classifications and only 1% of classifications are IPC 1. Given that we do not have many classifications in IPC 1 or any in IPC 5, we caution that our findings should not be applied to IPC 1 or 5.

Figure 1. Distribution of consensus-based phase classifications

![Figure 1. Distribution of consensus-based phase classifications](image)

This figure presents the relative frequency of each phase classification, with the actual counts (in parentheses), using the 1881 classifications that have corresponding information on underlying FSIs.

Given that the classifications use data on multiple FSIs, we next explore how often the FSIs concur in their suggested phase classification, and how often they differ. Figure 2 illustrates the Pearson correlation of the percent of the population in phase 3+ as suggested by the different FSIs across all of our observations where the following five FSIs are available (HDDS, FCS, HHS, rCSI, and LCS, a sample of 599 classifications). Note that the correlation of the percent of population in IPC Phase 3 or above (crisis or worse) across FSIs is relatively low, with the highest being 0.58 between HDDS and FCS, two indicators that are correlated by construction (since FCS is essentially a measure of dietary diversity, with weighted food groups).

Figure 3 illustrates the same Pearson correlation for a single country: Afghanistan, by TWG analysis (i.e., by round within a country). For the purposes of this study, we have the most rounds of TWG classifications for Afghanistan, allowing us to study TWG analyses over time.

We find that the correlation coefficients change substantially over time even within the same country. For example, the correlation between the percent of population in IPC Phase 3 or above (crisis or worse) assessed by FCS and rCSI varies from 0.41 in March of 2022, to 0.075 six months later. Similarly, the correlation of the percent of population in IPC Phase 3 or above (crisis or worse) for FCS and HDDS, which have the same constituent questions, varies between 0.94 in September 2021, changing to 0.089 in September of 2022, with March of 2021 and 2022 closer to 0.5.

The correlation results show that the populations identified by each FSI are not consistently correlated within space over time. This discordance in the underlying indicators underscores the challenges of reaching consensus and the importance of the consensus process.

While we chose Afghanistan because it has multiple TWG analyses, allowing us to examine trends over time, we also note that this period coincides with a highly dynamic situation, which could account for the lack of consistency. The country experienced major disruptions during the Islamic Republic’s return to power in the fall of 2021. The food security indicator HDDS was missing for 20 of 45 analysis areas in September 2022, which may have contributed to the observed differences between FCS and HDDS. Nonetheless, IPC operates in highly dynamic, prolonged, complex humanitarian crises. While the challenges Afghanistan faces are unique to that context, it is also the case that other countries would face similar a range of challenges regarding data collection timing, volatility, and other contextual factors. Thus, there are many plausible reasons for variation in TWG-led processes.
In this case, without access to the raw data, we are unable to discern whether the variation in correlations across time reflects data collection, cleaning and computation issues with the food security indicators (all of which occur outside of the IPC process) or if relationships among the populations in IPC Phase 3 or above (crisis or worse) based on FSIs actually vary over time within location. We return to this in our recommendations.

**Figure 2. Correlation Matrix: FSI implied 3+ population (%) for sample with All-5-FSI**

This figure presents a Pearson correlation matrix illustrating the relationships among the share of the population assessed to be in phase 3+ based on various FSI such as HHS, rCSI, FCS, LCS, and HDDS. The analysis specifically focuses on the sample of analyses that have all 5 FSIs (All-5-FSI sample; n = 599).
This figure presents a Pearson correlation matrix illustrating the relationships among the share of the population assessed to be in IPC 3+ based on various FSI such as HHS, rCSI, FCS, LCS, and HDDS. The analysis specifically focuses on the Afghanistan sample, and the results are presented by round. (n = 180)

Note: In the case of the (AFG: 2022-09) sample, 20 out of 45 analysis areas were lacking HDDS implied population estimates, leading to a notable difference in sample size compared to other time periods.
3.2.2. Reliability score

One possible explanation for the low correlation in the percent of population assessed to be in IPC Phase 3 or above (crisis or worse) by FSI is that some waves of data collection may be unreliable. Per the Technical Manual V3.1, IPC assigns reliability scores (R) to each FSI. These scores incorporate two components: the soundness of the method (M) and the time relevance (T) of the data. Both components are rated on a scale from 1 (limited) to 2 (good), followed by the alphabet identifiers M and T. The final reliability score (R) is determined by combining these components, with a score of R2 indicating the highest level of reliability and scores of R1- representing the lowest levels of reliability.

In practice, these reliability scores reflect the representativeness and timeliness of the sample (and do not refer to specific indicators captured in the survey). FSIs are generally collected in a single survey for each analysis area and all indicators from that survey appear to receive the same reliability score. Therefore, by construction, reliability scores do not vary across FSI within an analysis area.

In our sample, we find little variation in the reliability scores within analysis areas at a given point in time. That is, all FSIs within each area of analysis tend to have the same reliability score in a given TWG analysis. Further, reliability scores often do not vary within a country: we only observe 13 times out of 28 where reliability scores vary within a TWG analysis. Thus, we cannot use reliability scores to identify the reliability of a specific FSI in a classification.

![Figure 4: Variation in reliability score by country x round level analyses](image)

This figure displays the frequencies of the types of reliability score (left), along with its sub-components: time relevance (middle) and soundness of method (right), at the country x round level (n = 28 country x round level analyses).

Figure 4 illustrates the variation in reliability scores within data provided to a single analysis. There are four distinct cases. An analysis is classified as “Consistently High” if all regions in the period receive the highest reliability score (R2). High reliability is when data are collected from representative samples and in a timely manner. An analysis is labeled as “Consistently Low” if all regions in the period receive scores lower than R2. If there is variation in the reliability scores for an analysis, it is categorized as “Not Consistent”. Additionally, if the score is missing, it is denoted as “Not available”. Our analysis indicates that out of the 28 country x round level analyses included, 14 analyses (50%) demonstrate consistency in their reliability scores, either consistently high or consistently low.

3.2.3. Consensus-based phase classification outcome compared to FSI implied phase classification outcome

We next present summary statistics on the IPC phase classifications by the level of agreement in the underlying indicators. We begin by calculating the range in classifications suggested by the FSIs for a single observation (see Table 2). The first column (Range = 0) shows the percent of cases when all FSIs imply the same phase for that specific IPC analysis area in that specific time period. The second column (Range = 1) is the percent of cases when FSIs suggest phases that differ by 1 classification. For example, if for a specific IPC analysis area in Haiti in 2021, three FSIs suggest phase 2 and two FSIs suggest phase 3, that observation would be included in the Range = 1 column.

We find that the majority of observations have FSIs that imply phases that differ by 1. About 40% of observations that are deemed to be in phases 2 or 3 have underlying FSIs that imply phases of up to 2 levels of difference (e.g., in April 2021, there were several districts in Sudan where the rCSI suggested Phase 1 while the FCS pointed to Phase 3).
Table 2. Share of range in FSI implied classifications by consensus-based phase classification

<table>
<thead>
<tr>
<th>Consensus-Based Phase Classification</th>
<th>Range (Max-Min Difference) in FSI Implied Classifications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6.7%</td>
</tr>
<tr>
<td>2</td>
<td>6.5%</td>
</tr>
<tr>
<td>3</td>
<td>5.1%</td>
</tr>
<tr>
<td>4</td>
<td>2.9%</td>
</tr>
<tr>
<td>Total</td>
<td>97</td>
</tr>
</tbody>
</table>

This table presents a cross-tabulation comparing the range in FSI implied phase classification for a single IPC analysis area (columns) by the consensus-based phase classification made by each country x round level analysis (rows) (n=1881).

Note: Some of the indicators will not align with phase 1 and phase 5. For example, by construction, FCS never categorizes in Phase 1. Only HHS has specific cut-offs for Phase 5. The above analysis codes food security indicators that have a range of possible classifications as consistent with any classification in that range (e.g., rCSI score of greater than or equal to 19 can be phases 3-5. We treat such values as consistent with all three classifications).

In Table 3, the data in the second column show the percent of FSIs that are consistent with the observed consensus-based phase classification. We also observe that frequently the majority of the FSIs are not in agreement with the final analysis. For analysis areas deemed by consensus to be in phases 2 or 4 for example, the majority of FSIs imply the phase should be something else. Thus, in the majority of observations, there is not a single consensus phase suggested by a majority of FSIs.

Table 3. Share of consistent FSIs with consensus outcome and % of FSIs implying the majority class by consensus outcome

<table>
<thead>
<tr>
<th>Consensus-Based Phase Classification</th>
<th>% of FSIs Consistent with Consensus-Based Phase Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.77%</td>
</tr>
<tr>
<td>2</td>
<td>48.39%</td>
</tr>
<tr>
<td>3</td>
<td>52.12%</td>
</tr>
<tr>
<td>4</td>
<td>44.33%</td>
</tr>
</tbody>
</table>

In this table, the column labeled ”% of FSIs consistent with consensus-based phase classification” presents the share of FSIs in agreement with the consensus-based phase classification (n = 1881).

Note: 24% of the 1881 cases analyzed (452 cases) do not have a single mode, indicating significant variation among the FSI implied outcomes. To address this, we have assigned a new share for these cases. This is done by dividing the total number of FSI available by the number of unique modes within each classification. For example, if a classification with 4 indicators implies values of 2, 2, 4, and 4, and it exhibits two modes (2 and 4), while another classification with the same set of FSIs implies values of 2, 3, and 4 and has a single mode (3), we assign a share of 50% to both cases. Some indicators should not be used for classifying phase 1 and or phase 5. For example, FCS (by construction) never categorizes areas as Phase 1. The above analysis codes food security indicators that have a range of possible classifications as consistent with any classification in that range (e.g., rCSI score of greater than or equal to 19 can be phases 3-5. We treat such values as consistent with all three classifications).

These statistics suggest both a need for a consensus process - in that in the vast majority of cases, the underlying FSIs are not in agreement - instead pointing to more than a single phase for an IPC analysis area. In addition, the statistics suggest that the consensus process often generates results that diverge from the phase suggested by the majority of the underlying FSIs.
The next two figures illustrate the data from Table 2 and the second column from Table 3.

Figure 5: Distribution of range in FSI implied phase classification

![Figure 5: Distribution of range in FSI implied phase classification](image)

**Figure 5** demonstrates the range in FSI implied phase classifications. The range is computed as the difference between the implied highest FSI implied classification minus the lowest FSI implied classification. The x-axis is the consensus-based phase classification. 0 indicates perfect consistency, meaning all food security indicators point to the same phase classification (n = 1881).

The fact that few observations have all FSIs implying the same phase classification is evidence of the challenges faced by TWGs in the consensus process. While there are very few zeros, which would indicate consistency across FSI implied phase classifications, it is very common for TWGs to face situations where FSI-implied phase classification varies by one or two classifications.
Figure 6 illustrates the distribution of phase classifications, specifically in relation to the share of FSI that aligns with the consensus-based phase classification. The x-axis represents four distinct consensus-based phase classifications, while the y-axis represents the relative frequency of each occurrence (the share of FSI consistent with consensus-based phase classification), ranging from 0 to 1 with intervals of 0.2. A value of 1 indicates perfect consistency, implying that all FSI implied classifications match the consensus-based phase classifications (n=1881).

Similarly, Figure 6 illustrates the degree of consistency between FSI implied phase classifications and the consensus-based phase classifications. For phases 2 to 4, most FSIs are not in alignment with the consensus outcome, highlighting the substantial deviation between the consensus-based phase classifications and the underlying FSI implied phase classifications.
4. AFI RESULTS

4.1 Results for current status AFI

We begin by comparing the phase implied by different FSIs to the IPC outcome for each IPC analysis area in each country and round. In Figure 7, the vertical axis represents the proportion of IPC analysis areas where the resulting IPC phase categorizations were the same (gray), lower (blue) or higher (red) than the phases implied by the FSI.

Figure 7 demonstrates the difference between the consensus-based phase classification and the FSI-implied phase classification. The categories indicate the size of the difference. Gray = 0 indicates no difference between ‘country x round’ level phase classification and the FSI implied phase classification. As colors become darker blue, the consensus-based phase classification is lower than the FSI implied phase classification. As colors become darker red, the classification is higher than what is implied by the FSI. The x-axis is the country by (x) round. The y-axis indicates share of classifications in each category for each analysis (n=1881).
As can be seen from figure 7, implied classifications based only on the FCS are generally less conservative, i.e., they suggest a higher phase, than the resulting IPC classification (the bars are largely gray and blue). Conversely, the classifications implied only by the rCSI are generally more conservative than TWG actual classifications (i.e., the graph is largely gray and red). We also see substantial variation in the difference between the consensus-based classification and the classified implied by individual FSI across both country and round. Results from a similar analysis for the other three FSIs – HDDS, HHS, and LCS are presented in Annex 3, Figure A14.

### 4.1.1 Arithmetic (simple) mean classification estimates compared to consensus-based classifications

The above figures show variation in the classifications implied by different FSIs. One naive approach would be to average these, applying equal weights across all FSIs. As described above, the arithmetic mean is the simple average of the population classified in acute need (IPC 3+) for each food security outcome indicator. The observed IPC 3+ is the observed population requiring urgent action: the sum of reported populations in IPC 3, IPC 4, and IPC 5. Similar estimates can be computed for each food security indicator. The benefit of the arithmetic (simple) mean is that it offers a simple way to deal with competing values.

The findings above (Figure 2 and 3) showed that the correlations across food security outcome indicators are often weakly correlated, and differently correlated across TWG analyses (and therefore across space and time). In our findings below, we show the variation in populations classified according to each food security outcome indicator. In other words, different food security indicators identify different sized populations in IPC Phase 3 or above (crisis or worse). This underscores the challenge of classification - TWGs reach consensus on the IPC phase and classification drawing on diverse data. By graphing each food security outcome indicator’s 3+ population compared to the final, observed IPC classification, we can see the implicit weights on the different indicators in the consensus classification outcome.

As discussed above, while taking the arithmetic mean runs contrary to IPC guidance, the arithmetic mean serves as a point of reference and can help to identify the role consensus plays in population classifications. The average should not be interpreted as more or less accurate than TWG-led analysis classifications of 3+. However, it supports the identification of patterns, and shows how TWGs weigh different indicators in practice.

We present results for the sample of TWG classifications with the following five food security outcome indicators: HHS, rCSI, FCS, LCS, and HDDS. The total sample is 695 classifications across six countries: Afghanistan, Democratic Republic of Congo, Haiti, Madagascar, South Sudan, and Yemen.

The black dotted line depicts the hypothetical case where the observed consensus of the observed population in IPC 3+ (on the x-axis) corresponds perfectly with the share of population implied to be in IPC 3+ by the food security indicators (on the left-hand y-axis). The black line is the population in IPC 3+ based on the arithmetic mean (simple average) of populations meeting 3+ classification across the five indicators. The thick black line should be read using the x and the left-hand y axis.
The graph presents the 3+ population (%) classified for all analyses with five food security indicators (n = 695). The x-axis represents the share of the population the TWG assesses to be in phase 3+ through the consensus process. The y-axis represents the share of the population who would be assigned to IPC 3+ determined only by the food security indicator data (i.e., the population implied by each food security indicator). The 45-degree line depicts the hypothetical situation where the phase classification from the consensus share in 3+ is equal to the share implied to be 3+ by the food security indicators. The thick black line represents the arithmetic mean of the population implied to be 3+ by food security indicators. The gray shaded area shows the distribution of the cases classified and should be read against the right-hand y-axis.

**Note:** The All-5-FSI group (n = 695) includes analyses with a majority of classifications that include all five food security indicators. Within these TWG analyses, there are some classifications with fewer FSIs. The average line plot, specifically labeled as 'All 5 Indicator Averaged 3+', is based on a subset of the sample that includes only 599 classifications where all five FSI indicators are present.

The horizontal gap between the black line and the dotted line can be interpreted as the unobservable consensus process. When the thick black line is above the 45-degree line, the 3+ population implied by averaging across the food insecurity indicators is higher than the actual 3+ population reached through consensus. When it is below, the 3+ population implied by averaging across the food insecurity indicators is lower than the actual 3+ population reached through consensus.

For example, when IPC 3+ (or consensus-based 3+ population) equals 20% population, the arithmetic mean population in IPC Phase 3 or above (crisis or worse), defined by the five FSIs, is about 35%. In this case, the IPC consensus process resulted in fewer people being classified as in IPC Phase 3 or above (crisis or worse) than a simple average across the populations implied by the five food security indicators. The remaining lines illustrate the population in need of urgent action (IPC 3+) implied by the individual food security outcome indicators and should be read using the left-hand axis. The shading around each food security outcome indicator line is the 95% confidence interval.
The gray shaded area is the frequency of cases classified as IPC 3+ and should be read against the right-hand y-axis. For example, an IPC 3+ classification for 45% of the population is the most common case, happening about 91 times out of 695. The majority of classifications find between 20% and 60% of the population is in IPC Phase 3 or above (crisis or worse) (IPC 3+).

We observe variation in the populations classified across FSIs. The rCSI provides, on average, a lower estimate of the food insecure population relative to FCS and LCS. The HDDS and HHS most closely approximate the averaged food security outcome, as seen by their close tracking of the thick black line.

Further, as more of the population is in IPC Phase 3 or above (crisis or worse), the averaged (thick black) line more closely tracks with the consensus (dotted) line, suggesting that the consensus approach to classifying more of the population more closely approximates taking an average than when classifying a smaller population in need. There are fewer classifications, and therefore more noise at the extremes across all indicators.

In sum, the percent of population identified as in IPC Phase 3 or above (crisis or worse) differs across food security indicators. The IPC 3+ consensus outcomes fall somewhere in the middle of the food security indicators, suggesting that TWGs, on average, don’t tend to pick one (or two) preferred indicator(s) nor do they simply average across available indicators to reach consensus outcomes. Further, the average across food security indicators tends to be higher than the observed IPC 3+ outcome, particularly when fewer people are classified in IPC Phase 3 or above (crisis or worse).

Annex Figures A3 and A4 presents results similar to Figure 8 for other combinations of food security indicators (e.g., FCS, LCS, rCSI, and HHS) and for IPC 4+. Results are consistent with the sample with all five indicators. We note that comparing across country-specific figures (see the Annex Figure A2 with Afghanistan and Yemen), we find somewhat different relationships across food security indicators. We return to this in our regression results.

4.1.2 Arithmetic (simple) mean population estimates (%) compared to consensus-based population (%)

We next examine whether the IPC is generally classifying fewer people in need than an averaging process.

Figures 9 and 10 present the population classified as in phase 3 or higher (Figure 9) or phase 4 or higher (Figure 10) compared to the result of the arithmetic mean of the population classified in these same phases using the food security indicators. As can be seen, the distributions are not centered at 0, and we observe a general tendency of the IPC to underestimate the percent of the population at 3+ or 4+ compared to the FSI average. We note that this imbalance reaches through to the distribution’s tails, where the difference in the population in IPC Phase 3 or above (crisis or worse) can be quite large. For example, more than 15 of the observations have the IPC designating 40% less population in IPC Phase 3 or above (crisis or worse) than are suggested by the mean of the FSIs, while no observations have 40% or higher population designated as 3+ by the mean of the FSIs than designated by the IPC.
The histogram presents the frequency of classifications by the observed difference in the consensus-based 3+ population and the arithmetic mean of the FSI implied 3+ population. Blue bars represent classifications that assigned a smaller population as experiencing acute food insecurity (IPC 3+) than the population according to the simple average of the FSI implied 3+ population (%). That is, they assessed fewer people at 3+ than what is implied by a simple average of the FSI. The green bar depicts IPC population analyses that are consistent with the arithmetic mean. Red bars depict IPC population analyses that are less conservative than the simple average - that is, where the IPC assessed that a larger population was in phase 3+ than implied by the simple average of the FSI. Sample includes all classifications, regardless of the number of FSI available (n=1881).

Thus, on average, the TWG consensus process identified a smaller population in IPC 3+ food insecure than taking the simple mean of the population estimates across food security indicators.
The histogram presents the frequency of classifications by the observed difference in the consensus-based 4+ population and the arithmetic mean of the FSI implied 4+ population. Blue bars represent classifications that assigned a smaller population as experiencing acute food insecurity (IPC 4+) than the population according to the simple average of the FSI implied 4+ population (%). That is, they assessed fewer people at 4+ than implied by a simple average of the FSI. The green bar depicts IPC population analyses that are consistent with the arithmetic mean. Red bars depict IPC population analyses that are less conservative than the simple average - that is, where the IPC assessed that a larger population was in phase 4+ than implied by the simple average of the FSI. Sample includes all analyses, regardless of the number of FSI available (n=1881).

As in Figure 9, in Figure 10, we again see that consensus-based 4+ population (%) is most often less than the simple average based on the FSIs. We should note that cut-offs for some food security indicators are not available for higher phases (referred to as Non-Defining Characteristics). For example, among our sample of indicators, only HHS has cut-offs to differentiate between 4 and 5. While the IPC AFI Reference Table indicates that rCSI cannot differentiate between 3 and 4 (and 5) and should not be used to classify populations in IPC 3+ vs. 4+, in practice there is an informal rCSI cutoff score of 43 to differentiate between IPC 3 and IPC 4+. We use this informal rCSI cutoff in Afghanistan, Ethiopia, and Haiti, which have non-zero 4+ populations.

The Annex includes consensus-based 3+ population (%) by country x round level analysis in our sample. We see most countries classify a smaller population through consensus relative to taking an average of the populations based on the food security indicators. Figure A5 shows that of the fifteen countries, Haiti and Yemen IPC classifications include a higher population than the arithmetic mean; Sudan’s consensus classifications are slightly higher than the arithmetic mean population. Turning to specific country x round level analyses in Figure A6, we see that in Haiti results reflect a specific TWG analysis rather than a country-level phenomenon. The September 2022 Haiti TWG analysis over-classifies relative to the mean of the FSI implied 3+ population, but the September 2021 TWG analysis did not. We only have data on one Yemen TWG analysis (October 2020). Investigating if this relationship holds for other TWG analyses in Yemen could provide valuable insights.

Thus, we do not see evidence of systematically high consensus-based population (%) relative to the mean of available food security indicators. In country x round level analyses where this occurs, this could likely reflect dynamic moments (e.g., rapidly changing situation in Haiti in the fall of 2022) that were not captured when the food security indicators were collected.

Finally, we caution that in the above figures, consensus-based populations (%) are compared to an arithmetic mean of the populations suggested by the FSIs. In practice, TWGs are cautioned against using the simple average of FSIs. As they converge the evidence, they weigh different indicators differently.
4.1.3 Logit weights across data availability and severity

Our results thus far show that the consensus process involves more than simple averaging across food security indicators. Further, as shown in Annex figures, TWGs vary the degree to which they rely on different food security indicators. We use a logit model to estimate the magnitude of the relative weights on the different FSIs used by the TWGs in their classifications. We show that these FSI weights vary over time and space and over the degree of severity of the phase classification. Moreover, the weights vary by whether the problem is one of phase classification or population allocation.

Figure 11 presents the marginal effects from four logit models. All models have the dependent variable $y = 1$ for analysis areas with IPC Phase 3+, with $y = 0$ otherwise. The right-hand side (independent) variables include the population shares at 3+. Each quadrant of the figure presents the results for a different FSI data availability group. Comparing within a quadrant provides evidence about relative weighting of indicators within a given FSI data availability group. Comparing across the quadrants provides insight into how data availability is associated with changes in the weighting of the FSI indicators.

Within each group, weights on indicators largely are not statistically significantly different from one another, with the exception of HDDS receiving lower weight than most other indicators in group 1 and 3, and FCS receiving a higher weight than other indicators in groups 3 and 4. The confidence intervals around each FSI indicate that some FSI (e.g., rCSI) exhibit a large range of effects on consensus than other FSI with smaller confidence intervals (e.g., LCS). Weights seem to vary by location and TWG analysis, which we further investigate below.

Results also show clear differences in weighting (reliance on different indicators) depending on the subset of available FSI. We find that weights on each FSI change from group to group (across panels). For example, Group 1 analyses do not rely much on HDDS but weigh other indicators similarly. However, within Group 3, which includes FCS, rCSI, LCS, and HDDS, indicators are weighted very differently. In Group 3, HDDS is not statistically significantly different from zero, indicating that on average HDDS does not contribute information meaningfully different from zero when TWGs also have FCS, rCSI, and LCS. Note that the large confidence interval for rCSI (Group 2) implies its weight varies substantially between location and round even within the subset of available FSI.

Figure 11: Logit model marginal effects (or weights) by combinations of available FSI: 3+ phase classification

The figure presents the marginal effects from four logit models analyzing the consensus-based 3+ classifications. Independent variables include the population shares at 3+ for all FSI data available to the TWG. The four quadrants depict the marginal effects for each FSI for each data availability group.

Note: Each model is specified as $y = f(\text{FSI}3+\% + \text{country FE})$, where $y = 1$ if IPC Phase is 3 or higher, and 0 otherwise.
Figure 12 presents the marginal effects of the same analysis (logit models) for the 4+ population figures. As before, the right-hand side (independent) variables include the population shares at 4+ and the analysis is conducted separately for different FSI data availability groups. Comparing results in Figure 11 with Figure 12 suggests that the implicit weights on the FSI indicators change when the severity changes. In particular, TWGs appear to put considerably more weight on rCSI for 4+ than 3+ classifications and considerably less emphasis on HHS, which (right hand panel of Figure 12) is not significantly different from zero. Down-weighting HHS at higher levels of severity is not consistent with guidance from the AFI Reference Table, which shows that HHS is the preferred indicator to differentiate among 4+ phases. As in Figure 11, the rCSI spread is quite large, evidence that rCSI is weighted differently across TWGs.

Figure 12: Logit model marginal effects (or weights) by combinations of available FSI: 4+ phase classification

The figure presents the marginal effects from four logit models analyzing the consensus-based 4+ classifications. Independent variables include the population shares at 4+ for all FSI data available to the TWG. The two quadrants depict the marginal effects for each FSI for Group 1 on the top (n=599) and Group 2 (n=360) on the bottom. Groups 3 and 4 have too few observations and the models do not converge.

Note: Each model is specified as $y = f(FSI\ 4+\ (%) + \text{country FE})$, where $y = 1$ if IPC Phase is 4 or higher, and 0 otherwise.

Figure 13 presents results from the second panel of Figure 11 and the second panel of Figure 12 on the same graph for FSI data availability Group 2 to facilitate comparison of the relative weights across severity class within a FSI data availability group. We observe that the marginal effects of an increase in the percent of the population predicted to be in phase 3+ has a smaller effect on a location being in phase 3 or higher than the marginal effect of an increase in the percent of population predicted to be in phase 4+ has on the analysis area being in phase 4 or higher.

In other words, for most FSIs, the sensitivity of phase classifications to increases in the population appears higher at higher levels of severity. This result may be intuitive in that higher levels of severity are less likely to be observed with error and are therefore deemed to be more informative of true depth of need. The only exception is HHS which has a smaller marginal effect for phase 4+, which, being the preferred FSI to differentiate between phase 4 and 5 is surprising.
The figure presents marginal effects from the second panel of Figure 11 and the second panel of Figure 12 (both Group 2) on the same graph to facilitate comparison of the relative weights across severity class (n=360).

**Note:** 3+ and 4+ models are specified as Phase 3+ Dummy = f(FSI Implied 3+(%)+ country FE) and Phase 4+ = f(FSI Implied 4+(%)+ country FE), respectively.

### 4.1.4 OLS Weights on population estimates

We also analyze the relative FSI weights for the IPC population estimates. This approach offers more nuance in capturing food security severity relative to relying on the logit-based model of classifications alone, given the variation in population severity within classifications (IPC 2021).

Figure 14 uses these computed FSI weights (see Table A7 in Annex 3) from ordinary least squares (OLS) regression models at the TWG analysis level to predict population in IPC 3+. We show in the prior simple mean figures (Figure 8) the differences across the food security indicators and the consensus outcomes indicate that TWG analyses differ from taking a simple average. Figure 14 also includes a regression line (in purple) estimated using cross-country FSI-indicator weights. Regression allows us to better estimate the weighted combination of FSI used across TWGs than the simple mean. Compared to the simple averaged line, we see that the predictive line is closer to the consensus outcome (the dotted line). We find that prediction does well for populations in IPC 3+ that are most commonly observed, i.e., between 30-60%. The fact that the estimated slope is lower than the simple average of all five indicators (the black line) is further evidence of TWG conservatism on average.

We find that the predicted OLS line (purple line) tends to classify more populations in IPC 3+ relative to the consensus outcomes for lower population shares in IPC 3+. In contrast, the consensus outcomes classify a larger share of the population in IPC 3+ in more extreme cases of need. However, tails represent rare events for the OLS model, meaning fit will be poorer and should be interpreted with caution; this is similar for TWGs, which also rarely see cases with primarily tail outcomes (e.g., low fractions of populations in 3+ or high populations in 3+s).
The graph compares the share of the population in IPC Phase 3 or above (crisis or worse) (“IPC 3+”) identified by the consensus process (represented by a dotted gray line) to the share of the population in IPC Phase 3 or above (crisis or worse) identified by weighted and unweighted averages (represented by black and purple lines respectively) using classifications with All-5 FSI. The x-axis indicates the share of the population assessed to be in IPC 3+ through the consensus process, while the y-axis represents the share of the population assigned to IPC 3+ based on the food security indicator data with different weighting methods. The shaded areas around the lines show the confidence interval, while the gray shaded area represents the distribution of the classifications and should be interpreted in relation to the right-hand y-axis (n = 599).

4.1.5 OLS weights across TWG analysis rounds within countries

Finally, we analyze the three countries that have multiple rounds of TWG-level data to study variation in food security indicator weighting across TWG analyses within countries. We use OLS regressions to estimate IPC population results. Figures 15, 16, and 17 present these results for Afghanistan, CAR, and Madagascar. Our findings are as follows:

• It does not appear that TWGs within countries have consistently ‘favored’ (or ‘disfavored’) FSI metrics.

• Instead, results are strong evidence that TWGs within countries place different (implicit) weights on the same indicators across analyses. Thus, the consensus process does not appear consistent in terms of the aggregation of indicators across time.

• This variation over time could be strong evidence of consensus at work, could indicate that some indicators are missing, could indicate data collection challenges at a particular time (e.g., see Figure 3 on Afghanistan, discussed in 3.2.1).
• However, if certain indicators are not relevant in a particular country, we might expect that TWGs within that country would consistently downplay them across all analyses. For example, if LCS does not reflect the underlying nature of food insecurity well in a particular country, we might expect that LCS has a consistently lower weight across TWG analyses for that country. We do not see that degree of consistency of weights across TWG analyses within a country in the limited cases available.

We return to the implications of the second point in our discussion and recommendations. A lack of consistency in how TWGs use food security indicators over time within the same country opens up an opportunity to better document when TWGs consider certain indicators to be less useful than other indicators.

Figure 15: Coefficients on FSIs for consensus-based 3+ population by round: Afghanistan

Figure 15 presents the coefficients from four ordinary least squares (OLS) models regressing the consensus-based population in 3+ on the share of the population in 3+ according to each of the five FSI indicators. Each regression is run for each round of analysis conducted in Afghanistan (in March 2021, September 2021, March 2022, and September 2022). Lines represent the 95% confidence interval. Coefficients can be compared to zero, across FSI indicators for a given analysis, and across analyses for a given FSI (n=180).

Note: Each regression model is specified as \( y = f(FSI_{3+} \%) \) where \( y \) is consensus-based 3+ population (%)
Figure 16: Coefficients on FSIs for consensus-based 3+ population by round: Central African Republic

Figure 16 presents the coefficients from three ordinary least squares (OLS) models regressing the consensus-based population in 3+ on the share of the population in 3+ according to each of the five FSI indicators. Each regression is run for each round of analysis conducted in Central African Republic (in April 2021, September 2021, and September 2022). Lines represent the 95% confidence interval. Coefficients be compared to zero, across FSI indicators for a given analysis, and across analyses for a given FSI (n = 191).

Note: Each regression model is specified as \( y = f(\text{FSI 3+ (%)} \) where \( y \) is consensus-based 3+ population (%)

Figure 17: Coefficients on FSIs for consensus-based 3+ population by round: Madagascar

Figure 17 presents the coefficients from three ordinary least squares (OLS) models regressing the consensus-based population in 3+ on the share of the population in 3+ according to each of the five FSI indicators. Each regression is run for each round of analysis conducted in Madagascar (in May 2021, April 2022, and November 2022). Lines represent the 95% confidence interval. Coefficients be compared to zero, across FSI indicators for a given analysis, and across analyses for a given FSI (n = 47).

Note: Each regression model is specified as \( y = f(\text{FSI 3+ (%)} \) where \( y \) is consensus-based 3+ population (%)

4.1.6 Distributional analysis: “bunching”

Next, we pivot from considering the influence of different input data on the consensus outcomes and explore the distribution of the consensus outcomes themselves. As in the above section, we focus on the percent of population determined to be in IPC Phase 3 or above (crisis or worse) (IPC 3+). Figure 18 presents the frequency distribution of the percent of population in IPC Phase 3 or above (crisis or worse) and the phase classification resulting from the consensus outcome by IPC analysis area and time. The observed distribution of the number of observations at different levels of the % of population designated as IPC 3+ is not smooth (Figure 18). We observe a sharp uptick in cases in phase 2 relative to 3 near the cutoff. We also observe a greater number of observations at 15% of the population in IPC 3+ than at 20%. Second, nearly 104 observations are deemed to have 20% of the population in IPC Phase 3 or above (crisis or worse) but are still designated as Phase 2.\(^3\)

\(^3\) IPC GSU identifies this is an error at analysis level, speculating it could be that the TWG teams either made a mistake or were unsure how to classify; whatever the cause, this classification is not supported by IPC technical guidelines. Automated checks (discussed in Recommendations section below) could flag these in real time.
Several possible explanations could account for this apparent bunching just below the phase 3 threshold. First, the bunching might be an artifact of the fact that some observations at the 20% cutoff are determined to be in Phase 2 by the consensus process, creating the visual impression of bunching. Second, we might expect that most of the population is in phase 2 - higher than phase 1 but not in IPC Phase 3 or above (crisis or worse). Third, as noted in the methods section, it could be that TWG members require more evidence to designate populations (or analysis areas) as phase 3, leading to some clustering below the threshold. Such caution could also explain the considerable number of observations at 20% (and that therefore meet the criteria for Phase 3) but still designated as Phase 2. The observed bunching below the threshold does suggest Working Groups may be cautious when designating a location Phase 3.

The wide dispersion of the share of population in IPC Phase 3 or above (crisis or worse) demonstrates the large degree of variation in severity within each IPC phase category. Bunching results presented in Figure 18 also show that populations classified as 3+ can vary from 20% all the way to 80% within the IPC 3 category. The classification alone (without the population numbers) masks significant heterogeneity in severity.

The graph presents the frequency distribution of consensus-based phase 2 and 3 classifications with a range of shares of 3+ population assessed by the TWG. Points far to the right are analysis areas with a large share of the population assessed as phase 3+. Points far to the left are analysis areas with a low share of the population assessed as being in phase 3+. Analysis areas assessed as phase 2 are in yellow, and phase 3 are in orange. The dotted line depicts the 0.20 threshold meant to guide TWGs to classify analysis areas as IPC 3 when the population assessed IPC 3 meets or exceeds 0.20 share. Points just to the left of the dotted line therefore are points assessed as an IPC 2 that had 3+ population shares that IPC TWG assessed as very close to 0.20. The sample is all TWG analysis areas assessed as IPC 2 (n=2522) or IPC 3 (n=3827) for a total sample size n=6349. (Refer to Annex 1.3 for a detailed description of the sampling procedure.).
The graph presents the frequency distribution of consensus-based phase 3 and 4 classifications with a range of shares of 4+ population assessed by the TWG. Points far to the right are analysis areas with a large share of the population assessed as being in phase 4+. Points far to the left are analysis areas with a low share of population assessed as being in 4+. Analysis areas assessed as in phase 4 are shaded in red, and analysis areas assessed as phase 3 are in orange. The dotted line depicts the 0.20 threshold meant to guide TWGs to classify analysis areas as IPC 4 when the population assessed IPC 4 meets or exceeds 0.20 share. Points just to the left of the dotted line therefore are points assessed as an IPC 3 that had 4+ population shares that IPC TWG assessed as very close to 0.20. The sample is all TWG analysis areas assessed as IPC 3 (n=647) or IPC 4 (n=3827) for a total sample size n=4474. Refer to Annex 2.1.3 for a detailed description of the sampling procedure.

We also conduct an analysis of bunching by country. We present the analysis in Figure 20. For each country, we estimate the predicted share of the population classified as being in IPC Phase 3 or above (crisis or worse) based on a weighted average of the levels identified by the underlying FSIs. The pink line illustrates this predicted share; we allow weights to vary at the TWG-level. We also illustrate the percent of the population in IPC 3+ by using the simple average of the FSIs, illustrated by the blue line. We compare these distributions to the percent of population in IPC 3+ identified by the consensus process, illustrated by the black line. These results suggest that the distribution of the percent population identified in IPC Phase 3 or above (crisis or worse) exhibits more ‘bunching’ away from the 20% cut-off than others, where the 20% cutoff is illustrated by the red vertical line. Specifically, we observe potential bunching in Lebanon, Guatemala and the Democratic Republic of Congo.
This graph compares the share of the population in phase 3 or above (crisis or worse) ("IPC 3+") identified by the consensus process (illustrated in black) to the share of population in IPC Phase 3 or above (crisis or worse) identified by weighted and unweighted average, illustrated in pink and blue respectively. The 20 percent population in IPC Phase 3 or above (crisis or worse) used as a cut-off for phase 3 is illustrated by a vertical red line. Weights are calculated by running an OLS regression for each TWG analysis of the share of people in IPC 3+ against the share of people in IPC 3+ indicated by each FSIs (n = 1881).
4.1.7 Residual analysis: are there favored analysis areas with non-random influence.

As noted in the methods, we estimate weights on FSIs by TWG analysis, and then explore how the percent of the population categorized in IPC Phase 3 or above (crisis or worse) by the consensus process varies from the predictions from these regressions. In other words, we ask how the percent of population in IPC 3+ generated by the consensus process varies from what would be predicted by a weighted average of the FSIs where the weights are determined by TWG and analysis round. We conduct this analysis for the three countries with multiple rounds of data: Afghanistan, Madagascar and South Sudan.

In the upper panel of Figure 21, we regress the absolute values against IPC analysis area fixed effects and plot the absolute value of the residuals from the analyses over four rounds for Afghanistan. This figure uses shading to show the degree of deviation in each analysis area between consensus-outcome and the prediction using weighted average of the FSI indicators. Lighter colors represent areas with little difference between the consensus outcomes and the weighted average predictions while darker colors represent large average differences over the four rounds (light green borders identify which regional fixed effects are significantly different from zero). These deviations suggest that the consensus process arrives at phase classifications for specific analysis areas that depart from a simple reading of the FSI data for that location. As can be observed, the largest deviations are in the provinces of Baghlan and Helmand.

In the lower panel of Figure 21, we plot the value of those same residuals, but retain their sign to capture whether the consensus process classifies a higher or lower percent of the population as being in IPC Phase 3 or above (crisis or worse) than the simple mean approach. This process can be used to identify if there are “favored” classification areas, i.e. analysis areas that are consistently identified as being in greater need than would be predicted by the underlying FSI data. Provinces in blue are those analysis areas where the consensus process generates a lower phase than the weighted average percent of population suggested by the FSIs, while areas in red are where the consensus process generates a higher phase than the weighted average FSI number. Despite having a high deviation between consensus and predicted outcomes, Helmand does not appear to have systematically higher or lower percent of population in IPC Phase 3 or above (crisis or worse) predicted by the consensus process than the weighted average FSI outcome. Thus, it appears to be a location where there are frequent deviations, but those deviations can go either way. On the other hand, Baghlan is a dark blue suggesting that the consensus-based classification is systematically lower than what would be expected from the local FSI data. As above, provinces bordered in green are those analysis areas where the difference in consensus versus average prediction is significantly different from zero.

In future analyses drawing on more rounds for more countries, one could identify whether specific IPC analysis areas are systematically identified as having higher or lower levels of food insecurity than predicted by regression models. Note that systematic deviations should not be taken as a failure of the consensus process. As noted in the methods discussion, if there are systematic issues with, say, the representativeness of data collection in a location, driven, for example, by large IDP or refugee settlements that may go over- or undercounted, these systematic deviations from the FSI predictions may be warranted.
Figure 21: Map of systematic variance (upper panel) and direction of deviation (lower panel) in population in consensus-based 3+ population (crisis or worse) (%): consensus-based 3+ populations vs. model predicted 3+ populations

Figure 21 (upper panel) displays a map of the fixed effects of the absolute values of the residuals obtained from the country x round level OLS for Afghanistan (see Figure 15). Lighter colors represent areas with little difference between the consensus outcomes and the model-predicted weighted average outcome while darker colors represent large average differences over the four rounds. The lower panel presents a map of the fixed effects from an OLS regression regressing the residuals from the country x round level OLS for Afghanistan (see Figure 15). Analysis areas with positive fixed effects are shaded in red, negative are shaded in blue. Red indicates that the consensus-based outcomes indicate a higher proportion of the population in IPC 3+ relative to the OLS predicted population in IPC 3+. Green boundaries from both upper panel and lower panel indicate the analysis areas that have a statistically significant difference in the location coefficients between the consensus-based population and the...
prediction. Pooled across rounds (n=180).

4.1.8 Noisy Input Data

We next examine what happens to population classifications when the underlying data are noisy. We would expect that the convergence process is more important and complex if the underlying FSIs diverge more in their implied phase classifications. Thus, we would expect that the residuals (i.e., the unexplained portion of variance in regression models) might be larger when there is more noise. To explore this, we first plot the relationship between the absolute difference in the percent of the population deemed to be in IPC Phase 3 or above (crisis or worse) by TWG analyses versus the average percent of the population in IPC Phase 3 or above (crisis or worse) implied by the FSIs against the “noise” in the FSI indicators, defined as the difference between the maximum and minimum percent of population in IPC 3+ implied by different FSIs. (Recall that the measure of ‘reliability’ of the FSI data do not vary sufficiently for us to use them in this form of analysis.) The further to the right on the horizontal axis, the greater the gap between the percent of population in IPC 3+ suggested by different FSIs. The higher an observation is on the vertical axis, the larger the divergence between the percent of the population in IPC 3+ deemed by the IPC versus the FSI average. To examine whether the IPC measures of population in IPC 3+ converge to the simple mean of the FSIs, we split the observations between those where the designated population is higher in the IPC than the simple mean (i.e. where the TWG analysis overpredicts relative to the simple mean of the FSIs, in red) versus where the population designated by the IPC to be in IPC 3+ is smaller than that predicted by the FSIs (i.e. the TWG analysis underpredicts relative to the simple mean, in blue). We also fit two trend lines to the data.

As we move right on the horizontal access, the noise in the FSI indicators increases, meaning the range of population estimates based on the FSI indicators increases. The red line indicates overprediction of the consensus-population estimates relative to the arithmetic mean. As noise increases, the analysis areas where we observe overprediction (relative to the mean) tend to converge to the arithmetic mean. That is, when data are noisier, TWG analyses overpredict population less (that is, in the case of TWG overprediction: the gap between the TWG analyses and the mean estimates closes). When consensus process underpredicts relative to the mean (blue line) and noise increases, the difference between population designated using the mean FSI and the population designated through consensus tends to be larger. In other words, the TWG analyses underpredict more in the presence of noise. That is, in the case of TWG underprediction: the gap between the TWG population analyses and the mean increases. This suggests that in both cases, the TWG analyses becomes more cautious (relative to the arithmetic (simple) mean of the FSIs) as the information from the FSIs diverges more.

Figure 22. Relationship between the difference in Maximum (MAX) and Minimum (MIN) FSI implied 3+ population (%) and the absolute difference between the consensus-based 3+ population (%) and the arithmetic mean of FSI implied population (%)

Figure 22 presents a scatter plot of the observed range in the FSI implied 3+ populations for a given country x round level analysis (x-axis) and the absolute value of the difference between the (observed) IPC 3+ population analysis and the arithmetic mean of the FSI implied 3+ populations (y-axis). The data available to the TWG gets “noisier” (in terms of range) as one moves to the right on the x axis. The “residual” in the prediction by the TWG (relative to the arithmetic mean) gets larger as one moves away from
the origin on the y axis \( n = 1881 \).

We might anticipate that consensus outcomes are more likely to diverge from an outcome predicted by regression when input data are noisy. For four rounds of TWG data in Afghanistan, we first estimate a regression of consensus-based populations in IPC Phase 3 or above (crisis or worse) on FSI implied populations in IPC Phase 3 or above (crisis or worse). To test this, we store residuals from country x round level regressions for Afghanistan. We then regress the absolute value of the residuals against three measures of the noisiness of the input data (the range of phases predicted by the FSIs and the percent of FSIs that predict the majority classification.

![Figure 23. Regression results from regressing the absolute value of residuals against possible sources of noise](image)

The y-axis represents the coefficients and standard errors obtained from the regression analysis, which regressed the absolute value of residuals from country x round level regressions (see §4.1.5.) using four rounds of Afghanistan data against specific sources of errors that are unique to each district and analysis. On the x-axis, we listed the variables included in the regression, such as the range in FSI implied Phase outcome, % of FSIs consistent with consensus-based phase outcome, HFA received total (%) \( n = 160 \).

**Note:** HFA Total (%) is determined by multiplying the percent of the population that received humanitarian aid by the fraction of caloric needs that the humanitarian assistance, particularly food aid, fulfills for that population. For instance, if in certain areas 50 percent of households fulfill 50 percent of their caloric needs through humanitarian food assistance, we would assign a value of 25 percent as the HFA total (%) for those areas.

Somewhat counterintuitively, in Figure 23, we observe that the larger the range predicted by the FSIs, the lower the absolute value of the residuals, although this result is not statistically significantly different from zero. In other words, when the input data are noisy, exactly when we might expect the consensus process to diverge from using these input data, the consensus process generates a result that’s closer to the weighted average outcome predicted by the FSI data.

Also surprising, we observe that the percent of the FSIs in the majority class has virtually no effect on the distance between the consensus and simple mean outcome. Thus, we do not observe that noise in the input data affects the variability of the consensus outcome. Last, as with the above correlations, we observe the result that the higher the HFA, the more the consensus-based process generates the same outcome as the weighted average of the FSIs. As noted above, this result is a bit surprising if HFA is employed as input data. For example, one might expect that observing a large quantity of food assistance might render the FSI data less relevant. If this were the case, one would expect that the consensus outcome might tend to diverge from the weighted average FSI outcome when large quantities of HFA are delivered.

Next, we conduct the same analysis on the residuals themselves, presented in Figure 24. We observe that the range has little effect on the outcome. The percent of FSIs in the majority class tends to be related to a higher consensus outcome, suggesting that greater certainty is positively related to higher consensus outcomes. This finding is consistent with Figure 21 above - that consensus outcomes tend to be more conservative when input data are noisy. Last, once one controls for these two other drivers, HFA has little effect on whether the consensus outcome is larger or smaller than the weighted
average FSI outcome.

Figure 24. Regression results from regressing direction of deviation against possible sources of noises

The $y$-axis represents the coefficients and standard errors obtained from the regression analysis, which regressed the absolute value of residuals from country $x$ round level regressions (see §4.1.5) using four rounds of Afghanistan data against specific sources of errors that are unique to each district and analysis. On the $x$-axis, we listed the variables included in the regression, such as the range in FSI implied Phase outcome, % of FSIs consistent with consensus-based phase outcome, HFA received total (%) ($n = 160$).

4.2 AFI projections

TWGs produce current status analyses and projections, usually for the following 1-2 quarters, equivalent to three to six months into the future. We can examine the probability that an IPC analysis area is projected to be in phase IPC 3+ (projected status or PS) aligns with the later realized current status classification analysis (RCSA). This comparison helps determine if projections tend to over or underestimate the classification for each analysis area relative to the resulting observed current status.

We present results as the share of analysis areas with a projection at $t-1$ (or $t$-hat) consistent with the realized current status analysis at $t$. For example, a TWG could meet in January to project classifications in April. The TWG then meets in April and produces a current status analysis for April onward. In this case, April is time $t$ and January is time $t-1$. We compute the share of analysis areas with the projected status higher, lower or the same as the realized current status. We limit our analysis of projections that had an accompanying observed current status analysis in the next period.

Findings are consistency checks rather than checks on accuracy because we are unable to verify why the observed current status classification at time $t$ differs from the projected classification made in $t-1$. A very unexpected weather outcome, conflict or other shock (e.g., Covid-19) could not be incorporated into even the most careful projection but would drive location to have a better or worse outcome.

Alternatively, a projection of a food crisis might generate an increase in food aid or other support that generates an improved realized outcome. To estimate why projections and realized outcomes differ would require quantifiable information about which assumptions made in the projection period were realized and whether those assumptions impacted food security as anticipated in the current status. We return to this point later in the recommendations §7.

In Figure 25, we illustrate the percent of observations where the projections over- vs underestimate the percent of population in IPC Phase 3 or above (crisis or worse) relative to the realized outcomes. Our findings suggest that projections generally seem to equally over- and underpredict the realized current status outcomes. If projections induce a response (in terms of aid), we would expect the projections to systematically overpredict relative to the current status classifications.

Evidence also indicates that there is less ‘bunching’ of classifications in projections and more in current status, suggesting that current categorizations are more conservative than projections.
Figure 25: AFI Realized Current Status Analysis Classifications (RCSA) at time (t) vs. Projected Status (PS)

The figure compares Projected Status (PS) and Realized Current Status Analysis (RCSA) for the sample TWG classifications with projection window overlapping with realized current status. The figure presents the percent (%) of each incidence (PS < RCSA, PS = RCSA, or PS > RCSA) based on unweighted normalized value counts from samples where we can match PS t-1 and RCSA t. Green indicates when the Projected Status classification for each case matches the Realized Current Status Analysis for that case. Blue indicates cases with the Projected Status below the Realized Current Status Analysis. Red indicates cases in which the Projected Status was above the Realized Current Status Analysis (n = 3068).

In Annex, we present results by country and country x round analyses in Figures A9 and A10, respectively. Across both country and country x round analyses figures, we see that most projections match with the realized current status classifications, as indicated by the green bar. No country consistently over- or under-projects food security classifications relative to the realized current status, suggesting a TWG's projections are generally consistent with the next TWG's analysis of the current status. However, there are cases when projections were significantly less severe (Uganda in March, 2022) or more severe (Mozambique in June, 2017) than the realized current status.
We initially aimed to analyze all available AMN data. Due to data limitations, we revised our objective to conducting three proof-of-concept analyses for two example countries. Our analysis of AMN data posed substantial challenges, including issues pertaining to data cleaning and inconsistencies between evidence reliability scores and evidence levels. We provide detailed documentation of these issues in Annex 2. For our two sample countries, we chose Kenya and South Sudan. With these two country case studies, we can demonstrate what could be learned with additional data. Results from these analyses, however, should not be extrapolated to the larger universe of AMN classifications.

In Kenya, we have data from three rounds of AMN classifications (2020, 2021, 2022), providing a sizable number of classifications and utilizing different data sources. In South Sudan, we have data from two time periods within a single year (2022), allowing us to examine both current and projected classifications.

In AMN analyses, TWGs classify GAM (global acute malnutrition) using WHZ (weight for height) and or MUAC (mid upper arm circumference) indicators. Within a country, TWGs often have different combinations of WHZ information, MUAC information and information on contributing factors. WHZ and MUAC evidence can be from representative surveys, from sentinel sites, from historical evidence (representative at the time of data collection), from screening data for MUAC, and from downscaled data from higher administrative level surveys for WHZ.

While the IPC Technical Manual V3.1 (2021) provides guidance on which indicators (WHZ or MUAC) to use for classification and population estimations when both are available, TWG teams must consider several factors. First, the TWG team generally chooses the indicator with the highest reliability score for the final classification of an area.

When both are similarly reliable, WHZ tends to be preferred, reflecting that WHZ is standardized across the world population unlike MUAC, which varies based on environmental factors, rates of stunting, body shape and other factors. When classifications using MUAC compared to WHZ differ by two or more phases, the preference is to rely on WHZ, unless MUAC is two or more phases higher. In that case, analysts converge evidence using contributing factors, generally increasing WHZ scores by 1 phase based on the MUAC evidence. The IPC Technical Manual V3.1 (2021) points out that “MUAC thresholds can only be used in conjunction with the other contextual information by taking into account the immediate causes of acute malnutrition and the locally understood relationship between MUAC and WHZ prevalence, and by using the convergence of evidence approach” (p. 39).

Second, for population estimates, IPC guidance recommends that country analysts to use both MUAC and WHZ for calculation of population estimates to compute a third indicator: combined GAM prevalence. Combined GAM prevalence is derived from a formula combining this information to calculate population estimates (see IPC 2021, p. 62) and reflects that MUAC prevalence and WHZ prevalence often do not identify the same children as malnourished within the same sample (Grellety and Golden 2016). Combined GAM prevalence aims to avoid double counting of children who have low WHZ and low MUAC. Countries, however, decide how to calculate population estimates. For example, South Sudan uses combined GAM while Kenya uses GAM based on WHZ for its population estimates.

Third, compared to AFI analyses, AMN TWGs work with a mix of different WHZ samples reflecting limited and infrequent collection of AMN data. A TWG may also have GAM prevalence based on WHZ collected from multiple evidence sources (e.g., from both representative surveys and historical data). When multiple forms of WHZ-based GAM information are available, guidance in the IPC Technical Manual V3.1 supports selecting one to generate classification estimates for an analysis area (i.e., in Step 4). A current status classification can be based on historical data for some analysis areas and on current sampled data in others. When using historical WHZ information, TWGs need to account seasonality or other temporally specific factors to adjust the historical WHZ upwards or downwards to reflect the current status.
Overall, guidance on making GAM indicator choices is clear and TWGs document their choices in AMN Worksheets, including whether they make upward or downward adjustments to available indicators. However, further documentation on which approach to population is being used, how contributing factors influence classifications, or the magnitude of adjustments (beyond direction) made by TWGs to account for adjustments to the GAM indicator (e.g., due to historical data) and why is somewhat limited. This means it is challenging for outside observers to understand how particular classifications and population estimates are arrived at.

Additional data-related challenges limit our analyses to proof-of-concepts. Comparing projections to current realized AMN analyses requires overlapping periods to examine whether the projections made for time t in time t-1 made match realized current status in time t. Our projection and realized current status data do not overlap. South Sudan has a projection for a window a few months prior to the current status. We use this as a proof-of-concept. However, acute malnutrition can be highly sensitive to seasonality and we underscore that the lack of overlap means the South Sudan example should not be treated as a concrete finding. Finally, we note that the reported AMN information is time intensive to collate and some is not machine readable, as described in Annex 2.

5.1 AMN proof of concept data and methods

Given these limitations, we focus on presenting proof-of-concept analyses. We first evaluate the data, its sources, timeliness, and representativeness through descriptive statistics for our two proof-of-concept countries. We then conduct three analyses using AMN consensus outcomes and projections to:

1. Evaluate the evidence level available to TWGs.
2. Evaluate the consistency between representative and non-representative survey-based outcomes.
3. Evaluate the consistency between implied classifications based on WHZ and MUAC indicators and the Consensus Classifications reached.
4. Evaluate the consistency between projections made for time t in t-1 and current status at time t.

We first examine the consistency of findings compared to the IPC AMN Reference Table based on their data sources (i.e., historical evidence, representative samples, or other), and other features of AMN data.

Second, we explore the consistency of TWG observed classifications with the underlying nutrition data. We use the AMN reference table provided in the IPC Technical Manual V3.1 (2021), to ascertain the “mechanical” classification for each available indicator (i.e., without consensus) and compare this to the consensus outcome. We show results for all available indicators, although in practice TWGs often prioritize a specific indicator to use for classification. At scale, such an analysis would help us understand how information on contributing factors influence consensus classifications.

Third, we demonstrate whether projections made for time t in time t-1 are consistent with current status in time t if overlapping data were available.

While these case studies give us insight into the consistency between the input data and the outcome in these specific settings, conducting this comparative analysis on a larger scale would yield valuable insights into the degree of inconsistency that may arise when using alternative data sources in the absence of representative survey data.
5.2 Descriptive statistics and level of evidence

We present summary statistics for Kenya and South Sudan in Table 4.

Table 4. Summary statistics for Kenya and South Sudan

<table>
<thead>
<tr>
<th></th>
<th>Kenya (2020, 2021, 2022)</th>
<th>South Sudan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mar 2022</td>
<td>Oct 2022</td>
</tr>
<tr>
<td>Total Classifications</td>
<td>56</td>
<td>77</td>
</tr>
<tr>
<td>Used WHZ for classification</td>
<td>55</td>
<td>77</td>
</tr>
<tr>
<td><strong>Source of Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representative Survey Data</td>
<td>42</td>
<td>14</td>
</tr>
<tr>
<td>Higher Admin &amp; No Rep Survey Data (No Others)</td>
<td>0</td>
<td>38 (17)</td>
</tr>
<tr>
<td>FSNMS Data &amp; No Rep Survey Data (No Others)</td>
<td>0</td>
<td>10 (10)</td>
</tr>
<tr>
<td>Historical Trend &amp; No Rep Survey Data (No Others)</td>
<td>11</td>
<td>44 (11)</td>
</tr>
<tr>
<td>Similar Area &amp; No Rep Survey Data (No Others)</td>
<td>0</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Sentinel Sites &amp; No Rep Survey Data (No Others)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No Data</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Used MUAC for classification</strong></td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

This table provides a summary of the sample used for AMN evaluation. The sample consists of 203 AMN classifications from 5 TWG analyses.

The majority of consensus-based AMN phase classifications are based on the Weight-for-Height Z-scores (WHZ). When classifications are made with MUAC-based GAM, contributing factors must be included. Data for AMN classifications are sourced from multiple channels, including representative surveys, higher administrative data, FSNMS (Food Security and Nutrition Monitoring Systems) data, historical evidence, data from similar areas, and sentinel sites.

Of the 203 area classifications included in our analysis, TWGs classified 56% using representative surveys, while the remaining 44% relied on alternative data sources. In Kenya, 20% of the classifications were based on historical evidence, and in South Sudan (March 2022), this figure was 14%. In South Sudan, a representative survey was fielded between March and October of 2022, such that the AMN TWG had newer WHZ data to work with for 58 of 70 classifications in October.

AMN data have different evidence levels than those used for FSI data. Evidence level in AMN analyses is a technical term (see IPC Technical Manual V3.1 p. 166 for more details). For current status analyses, high evidence level (*** or level 3) indicates the availability of WHZ data from timely representative surveys. Medium evidence level (** or level 2) includes somewhat reliable WHZ or MUAC data, meaning the data collection was timely but may not be representative and includes at least two pieces of evidence on contributing factors. Acceptable evidence level (*) or level 1) includes representative WHZ data that is less timely (see Figures 131 and 132 in IPC Technical Manual V3.1 for more details) and includes at least two pieces of evidence on contributing factors.

The existence of WHZ data alone – or WHZ from representative surveys – does not automatically imply high evidence level designation, as other factors like timeliness and sample size can influence that designation. For example, in South Sudan in October 2022, almost half of the WHZ classifications are based on evidence labeled as level 1 (poor data reliability, small samples, or relying on sources other than representative survey from the ‘t’ period). Further, WHZ based classifications (n=65) rely both on representative survey data (n=56) and historical data (n=11) for classification.
Figure 26. Evidence level for Kenya and two time periods in South Sudan (Evidence level is a function of type of data, timeliness of collection, and representativeness).

The Kenya sample, pooled from 2020, 2021, and 2022, includes 56 cases. South Sudan in March 2022 includes 76 cases. South Sudan in October 2022 includes 70 cases. High evidence levels are denoted by *** and green color. Medium evidence levels are denoted by ** and orange color. Acceptable evidence levels are denoted by * and blue color. See Figures 131 and 132 in IPC Technical Manual V3.1 (2021) for more details on how evidence is classified based on type of data, timeliness of collection and representativeness.

Figure 26 shows the evidence levels. The Kenya outcomes for 2020, 2021, and 2022 show that the TWG had high (*** evidence levels for 32% of the cases. 48% of the Kenyan cases were classified as acceptable (*) evidence. In South Sudan in March and October of 2022, 5% and 20% respectively were high evidence classifications. Thus, it is common for AMN TWGs to classify based on evidence other than “high” and to need to draw on contributing factors to do so.

5.3 Consistency between representative and non-representative survey based AMN phase classifications

Figures 27 and 28 show that relying on non-representative data (i.e., historical evidence, higher administrative level data downscaled, FSNMS data, data collected from sentinel sites or data other similar analysis areas) for classifications increases the likelihood of the consensus outcome differing from the outcomes that reflect the classification indicated by the IPC AMN Reference Table values.

Figure 27. Consistency between WHZ implied AMN phase classification derived from the IPC AMN Reference Table and the consensus-based AMN phase classification by evidence type for Kenya

The Kenya sample, pooled from 2020, 2021, and 2022 classified using WHZ evidence (n = 56). 42 are analysis areas that were classified using representative WHZ data. 11 were analysis areas classified using historical WHZ evidence. Three cases, not presented here, were classified either from no underlying data, data from sentinel sites, or MUAC data. The blue color indicates cases where the implied classification derived from the AMN Reference Table is consistent with the actual TWG consensus. The orange and green color indicate cases where the classification derived from the AMN Reference Table is different from the classification reached through the actual analysis.
In Kenya, when using historical WHZ data, 36% of the TWG classifications differ from what the underlying data suggest (i.e., are either higher or lower). On the other hand, classifications based on WHZ data from representative surveys consistently match the phase suggested by the underlying data.

**Figure 28.** Consistency between the WHZ implied AMN phase classification derived from the IPC AMN Reference Table (bars) and the consensus-based AMN phase classification for South Sudan (x-axis) (March 2022)

The dataset used in this analysis includes 70 cases extracted from the South Sudan AMN worksheets (March 2022). The x-axis illustrates the consensus-based AMN phase outcomes. The y-axis is the frequency of classifications for WHZ input data. The panels represent the characteristics of the input data (WHZ REP: Representative Survey Data; WHZ HIGH ADMIN: Data from higher admin level; WHZ REP FSNMS: FSNMS Data; WHZ SIMILAR AREAS: Data from similar areas; WHZ HIST: Historical Data).
Figure 28 illustrates the WHZ implied AMN phase classification, and the consensus-based AMN phase classification by type of input data. When the underlying data are representative and come from either surveys or higher administrative data, downscaled by the TWG, the resulting consensus phase classifications largely match the phases implied by the underlying data. On the other hand, when the data are historical, the consensus outcomes tend to be different from what are suggested by the mechanical calculations associated with the underlying data. More than a quarter (27%) of phase 2 outcomes were suggested to be phase 3 and 13 percent were suggested to be phase 4. Similarly, 19% percent of phase 3 outcomes were suggested to be phase 2 and 38% were suggested to be phase 4. Most consensus-based phase 4 classifications were suggested to be phase 4 by the underlying data, while 20% were suggested to be lower.

These findings are not surprising, as TWG analyses are adjusted upwards or downwards historical information based on changes in contributing factors (e.g., immediate and underlying causes) to reach consensus.

5.4 Consistency of results among WHZ, MUAC and AMN consensus-based classifications:

In practice, TWGs often select the indicator (WHZ or MUAC) based on evidence reliability, which is then use for classification (IPC Technical Manual 2021).

In what follows, we examine the variation between the implied classifications based on available underlying indicators and the consensus classifications. In Figure 29, we explore the relationship between the WHZ implied outcome and the realized consensus-based phase outcome in Kenya. We use 42 out of 55 classifications that use GAM based on WHZ representative survey data, considered the most reliable for measuring acute malnutrition in the IPC Technical Manual V3.1. The figure shows that classifications implied from WHZ values from the representative survey align closely with consensus-based classifications.

The dataset used in this analysis is 42 cases extracted from the Kenya AMN worksheets (2020-2022). The included cases are those with Weight-for-Height Z-score (WHZ) measurements obtained from a representative survey. The x-axis is the consensus-based AMN phase classification. The y-axis is the frequency of WHZ implied phase classifications.
Similar trends are observed in Figure 30 for South Sudan, applicable to both March and October 2022 TWG analysis. The WHZ values based on representative survey implied phase outcomes align consistently with the consensus-based classifications (on the x-axis). However, the process is not entirely mechanical. For example, in March 2022 TWGs chose to adhere to higher administrative level data, indicating a phase one level lower. This difference between the implied outcomes and consensus outcomes shows that there are cases, even when utilizing the most reliable data available, when consensus outcomes diverge from the implied outcomes.

**Figure 30. Consistency between WHZ implied AMN phase classification (bars) and consensus-based AMN phase classification (x-axis) for South Sudan**

The dataset used in this analysis is 65 cases extracted from the South Sudan AMN worksheets (2022). The included cases are those with Weight-for-Height Z-score (WHZ) measurements obtained from a representative survey. The x-axis is the consensus-based AMN phase classification. The y-axis is the frequency of WHZ implied AMN phase classifications.
To summarize, for this small sample, we find that AMN TWG consensus-based outcomes are highly consistent with WHZ GAM implied classifications using the reference table. In other words, when WHZ representative data are available, the consensus-based classifications are generally consistent with what is implied by these data.

There are several avenues for future research with a larger sample. First, a larger sample size would let us speak to findings beyond these two country cases. Second, we could explore cases with both historical and representative WHZ to understand the influence of MUAC information on consensus-based outcomes. For example, we could examine whether a lower range or an upper range of MUAC is more likely to change the consensus classification. In other words, are TWGs likelier to place higher weight on MUAC values that push the classification up or that push the classification down?

5.5 Current status and projections AMN

Using an approach identical to examining consistency of the AFI projection analyses, we demonstrate an example comparing projected AMN status (PS) classifications with a realized current status analysis (RCSA). However, we underscore that we do not have any cases where the AMN projection window matches the current status window and therefore the following results are for demonstration purposes only.5

As a proof-of-concept we examine South Sudan projections, published in March 2022 and made for June to August 2022 against the nearest realized current status analysis results, which were observed in October 2022. No analysis was done in June to August, which was the projection window. June and July are generally periods of peak acute malnutrition in South Sudan, while acute malnutrition declines October onward. Thus, we expect that the projections differ significantly from the current status analysis several months later. Therefore, results should therefore be interpreted as a reflection of the accuracy of analysis in South Sudan.

Figure 31. Consistency of South Sudan projected status (PS) with nearest current realized status analyses (RCSA)

The figure displays the distribution of differences between realized current status (RCSA) outcomes in October 2022 and projected status (PS) in South Sudan in March 2022. The plot is color-coded, with blue indicating ‘PS < RCSA’, green indicating ‘PS = RCSA’, and red indicating ‘PS > RCSA’. Each column on the x-axis represents a data source used for the current status analysis in March 2022, including ‘FSNMS’, ‘Higher Admin’, ‘Historical Data’, and ‘Representative Survey’. The y-axis represents the relative frequencies of the cases.

Note: In the March 2022 sample, some classifications utilize historical data for projecting AMN levels in June to August 2022. However, the majority of the realized current status analyses in October 2022 are based on representative survey data.

5 A further challenge for AMN TWGs is that in some cases, the GAM current status at t-1 is based on historical data (see §5.2. desc stats). The projection is then based on already dated information in t-1 and is projected forward to time t. If new data are collected for the realized current status analysis at time t, we cannot isolate whether a mismatch between projections and current outcomes is because the projections are based on extrapolations from historical information that are unlikely to match updated information, the situation changed in unforeseen ways, or whether the projection process itself was inaccurate.
We categorized the data into four different types of sources based on the source of the data for the current status classification at time t-1 (or March 2022 in this specific example). These sources include Representative Survey with 14 samples, historical evidence (with dated representative survey data) with 13 samples, higher admin data (with no representative survey data) with 30 samples, and FSNMS data (with no representative survey data) with 9 samples. In this proof-of-concept, we observe that most projections are the same as the later consensus outcomes. In slightly more cases, projections are a higher phase than later realized consensus outcomes compared to cases where the projections are lower than realized consensus outcomes. Unlike the FSI data, we observe that the projections using historical data are, if anything, closer to the realized consensus outcomes than those using more representative data.

Future work with cases with projection windows matching realized current status analyses would allow us to discern patterns regarding first, whether projections under or overstate food insecurity relative to the relative current status and second, whether certain data types are more prone to these challenges. For example, in this example case, projections relying on FSNMS data seem to overstate acute malnutrition relative to other data forms (we underscore that projection and current status windows do not match and this should be treated as a proof-of-concept). Further, we note that it is to be expected that projections might be systematically above or below realized current status analyses due to unforeseen shocks.
To contextualize the findings that follow, we emphasize that our findings focus on outcome indicators analyzed by TWGs during the consensus process. The consensus process is both important and challenging. Food security and malnutrition are latent constructs with no measure of the underlying ‘true’ level of food insecurity. That is, as the challenge of developing this research has demonstrated, TWGs do not have an objective truth against which to compare phase outcomes. It also raises challenges to examining the accuracy of the IPC. Because we cannot explicitly measure the ‘accuracy’ of the phase or population classifications, we instead focus on exploring the consistency of the process (see §2.2). Given richness of AFI current status data and limited availability of AMN data, the majority of our discussion is drawn from the AFI current status findings.

We sought to answer four main questions.

- First, we ask “What do the underlying indicators suggest to TWGs about phase classification and population outcomes?” This question seeks to explore the information TWGs have.
- Second, we ask “What is the consensus process doing?” This allows us to examine if population or phase classification outcomes are systematically biased.
- Third, we ask “What could be influencing consensus outcomes?” This focuses on what could be shaping the outcomes we observe.
- Fourth, we ask “What is the relationship between projections and current status classifications?” This allows us to examine if projections are systematically higher or lower than consensus classifications.

### 6.1 AFI current status

We answer the first through third questions below. Question four is discussed below in the AFI projections section.

#### 1. What do the underlying indicators suggest to TWGs about phase classification and population outcomes?

The underlying FSI data vary substantially between TWGs, making consistency in phase classifications and the percent of population in IPC Phase 3 or above (crisis or worse) across TWG’s consensus processes difficult. The set of FSIs and their timeliness can vary substantially, making comparability challenging (see §3.2.1). The FSIs capture different aspects of food security and are only loosely correlated with each other (see §3.2.1). As a result, a TWG using one subset of FSI may capture a very different aspect of food insecurity than a TWG using a different subset. This difference in input data makes it difficult to compare the consensus process and outcomes across TWGs, even within the same country over time.

The consensus process is essential given that the underlying FSI data tend to suggest different phases, implying the appropriate phase classification is not obvious. This lack of consensus in the underlying data imply that using a simple average to examine the appropriate phase classification (or % of population) is generally not possible. Instead, the TWG must decide which FSI information to weigh more heavily in their analysis. We present descriptive statistics that show that in the majority of cases, the FSIs suggest different phase classifications for the same IPC analysis area (see §3.2.1), demonstrating the importance of the consensus process to determine the correct phase. For 43% of cases, phase levels implied by the FSI differ by 2 phase classifications or more (see Table 2 in §3.2.3). Further, while the reliability scores highlight that only 25% of FSI data are highly reliable, reliability measures largely do not vary by TWG, implying that the TWG has little external guidance on how to compare information from FSIs that point towards alternate phase classifications (see Figure 4 in §3.2.2).

#### 2. What is the consensus process doing?

We find that most of the available FSI data are used in the consensus outcome. When we regress consensus phase and population estimate outcomes against the FSI derived population levels, we find that most FSIs significantly contribute to the consensus outcomes. HDDS is a possible exception; HDDS is rarely statistically different from zero in regressions covering all TWG analyses that have HDDS as an FSI input. Similarly, few TWGs use LCS (see §4.1.3). This is consistent with findings from our qualitative interviews in which several respondents raised concerns about whether LCS appropriately measured food security in some settings.
We find evidence that the IPC is, on average, classifying fewer people as being in IPC Phase 3 or above (crisis or worse) (i.e., is conservative) relative to the FSI reference points. We first plot the percent of the population in IPC Phase 3 or above (crisis or worse) (IPC 3+) from the consensus outcome to the mean of the levels predicted by the FSIs. We see illustrative evidence that overall, the consensus measure tends to be lower for most of the distribution of outcomes than the mean (Figure 8).

We next formally compare the two measures: the IPC consensus outcomes for the percent of the population in IPC Phase 3 or above (crisis or worse) (IPC 3+) or in IPC 4+ compared to the mean levels predicted by the FSIs. We find that the consensus-based phase outcome is more often lower than the phase predicted by the FSIs for both IPC 3+ and IPC 4+ levels (see §4.1.1 and §4.1.2). We also compare the simple average of the FSIs’ predictions of the population in IPC 3+ to the weighted average, where the weights are estimated to predict the consensus population outcome. We find that the weighted outcome underpredicts the proportion of population in IPC Phase 3 or above (crisis or worse) compared to the simple average for all severity levels (§4.1.2). Based on interviews reported in the Scoping Study, we were expecting that TWG-led analysis might be more likely to over-classify populations.

We plot the distribution of the percent of population deemed to be in IPC Phase 3 or above (crisis or worse) by phase classification (§4.1.2). We find needs may be very different within the same phase classification. A single phase classification can encompass a wide range of percentages of population in IPC 3+. Specifically, analysis areas classified as phase 3 have percentages of population deemed to be in IPC 3+ ranging from 20 to as high as 80 percent. This finding underscores that while the phase (i.e., the color on the map) may look the same for two different analysis areas, the same classification can mean very different things with respect to the severity of the circumstances.

Next, we consider the distribution of the percent of the population in IPC Phase 3 or above (crisis or worse) versus the phase classifications. We find some evidence of ‘bunching’, a mass of the distribution just below the 20% threshold, that would suggest that a location is in phase 3. We observe several analysis areas where the population in IPC 3+ is above 20% but the analysis areas are categorized as phase 2. In analyses by country, we find several countries with a mass of observations just below the 20% threshold (see §4.1.6). This is suggestive evidence of caution by TWGs deciding to classify a location as phase 3, indicating that the degree of caution varies by TWG. Last, when we interrogate the deviations of convergence outcomes relative to the level of population in IPC Phase 3 or above (crisis or worse) predicted by a weighted average of FSI data, we find suggestive evidence that the deviations are larger when the FSIs predict higher phases, again suggesting the TWGs are taking a cautious approach (see Appendix §3.5).

There are at least two ways to interpret these results related to the conservatism of the TWGs. The results comparing consensus outcomes to the average FSI and illustrating the distribution of the population in IPC Phase 3 or above (crisis or worse) may reflect that some TWGs prefer a higher evidence quality level to meet a phase 3 classification (which indicates urgent need). As a result, TWGs may shift some potential 3 analysis areas to level 2. On the other hand, these results could be related to selection: life in a low-income country where IPC classifications happen is precarious, so there are few cases of IPC = 1; yet, many may not be in crisis (IPC = 3), so 2 is where a large proportion of the population lands. The fact that we observe that the convergence outcomes tend to be more conservative when input data are noisier, may lend support to the first hypothesis - that when input data are contradictory, a TWG may prove cautious to classify a location or a high proportion of the population as in phase 3.  We observe several analysis areas where the population in IPC 3+ is above 20% but the analysis areas are categorized as phase 2.  In analyses by country, we find several countries with a mass  needs may be very different within the same phase classification.  A single phase classification can encompass a wide range of percentages of population in IPC 3+. Specifically, analysis areas classified as phase 3 have percentages of population deemed to be in IPC 3+ ranging from 20 to as high as 80 percent. This finding underscores that while the phase (i.e., the color on the map) may look the same for two different analysis areas, the same classification can mean very different things with respect to the severity of the circumstances.

Next, we consider the distribution of the percent of the population in IPC Phase 3 or above (crisis or worse) versus the phase classifications. We find some evidence of ‘bunching’, a mass of the distribution just below the 20% threshold, that would suggest that a location is in phase 3. We observe several analysis areas where the population in IPC 3+ is above 20% but the analysis areas are categorized as phase 2. In analyses by country, we find several countries with a mass of observations just below the 20% threshold (see §4.1.6). This is suggestive evidence of caution by TWGs deciding to classify a location as phase 3, indicating that the degree of caution varies by TWG. Last, when we interrogate the deviations of convergence outcomes relative to the level of population in IPC Phase 3 or above (crisis or worse) predicted by a weighted average of FSI data, we find suggestive evidence that the deviations are larger when the FSIs predict higher phases, again suggesting the TWGs are taking a cautious approach (see Appendix §3.5).

There are at least two ways to interpret these results related to the conservatism of the TWGs. The results comparing consensus outcomes to the average FSI and illustrating the distribution of the population in IPC Phase 3 or above (crisis or worse) may reflect that some TWGs prefer a higher evidence quality level to meet a phase 3 classification (which indicates urgent need). As a result, TWGs may shift some potential 3 analysis areas to level 2. On the other hand, these results could be related to selection: life in a low-income country where IPC classifications happen is precarious, so there are few cases of IPC = 1; yet, many may not be in crisis (IPC = 3), so 2 is where a large proportion of the population lands. The fact that we observe that the convergence outcomes tend to be more conservative when input data are noisier, may lend support to the first hypothesis - that when input data are contradictory, a TWG may prove cautious to classify a location or a high proportion of the population as in phase 3.  Relatedly, we find that when input data are more consistent, the consensus outcome is more likely to be higher (see Annex §3.5).

3. What could be influencing consensus outcomes?

We examine the consistency of consensus outcomes across TWG analyses. In our regression models we find that weights on FSI vary over time and space. They also vary over the degree of severity of IPC classification. When we first plot the percent of the population in IPC Phase 3 or above (crisis or worse) from the consensus process to the percent of population in IPC Phase 3 or above (crisis or worse) (IPC 3+) suggested by each FSI, we find the expected result that the consensus outcomes appear to weigh each FSI differently (§4.1.1). HHS and HDDS track the consensus outcomes most closely, while rCSI underpredicts the consensus outcomes, and FCS and LCS both predict a larger percent of the population in IPC 3+. We then estimate the weights on each FSI for the classification and the estimation of the percent of classified population by (a) the set of available FSIs and (b) country and round. We also compare findings to estimated weights on each FSI for IPC Phase 4 or above outcomes. We show that weights vary substantially over all of these dimensions.

The fact that weights vary is not in and of itself an indication of an issue with the process. One can view these outcomes as either raising potential concern about inconsistency or as demonstrating the value of the consensus process – evidence for example that TWG members have a common language and approach and use their knowledge to adjust the data
to better match what they understand is happening on the ground. However, this finding suggests that it may be misleading to compare consensus outcomes across TWG analyses because our analysis suggests that each TWG analysis uses food security indicator information differently to achieve consensus, even within the same country over time. As we discuss in the recommendations section, additional documentation from the TWG regarding which indicators (and contributing information) are more heavily relied upon and why will help clarify comparability. Documentation can also support understanding of whether variation in weights reflect contextualization and consensus or if they reflect "loudest voices in the room" or other influential sources of bias. We also note that our analysis does not include contributing factors and including such information could improve our understanding of the consistency of findings across TWGs.

As noted in the methods section (§3), if one indicator is deemed to be less relevant in a particular country, one would still expect that indicator to be consistently (down) weighted in that TWG analysis. Instead, we see weights on different FSIs varying greatly within TWG analysis and within countries across rounds. This suggests that TWGs are not treating the input data the same way for each location. Our findings of large variation in weights on specific food security indicators suggests that using the IPC as an outcome variable is not equivalent to using another measure of food security, such as the reduced coping strategy index (rCSI) or food consumption score (FCS).

For a small sample of analysis areas where we observe multiple rounds (Afghanistan, Madagascar and South Sudan), we find the percent of population in IPC Phase 3 or above (crisis or worse) determined by the consensus process deviates significantly from the simple mean prediction in most analysis areas (§4.1.7). When we explore these deviations, we do not find strong evidence that some analysis areas are consistently rated to be in greater or less need than would be predicted by the FSI data. We compare the consensus outcomes to outcomes predicted by TWG-level OLS regression models. With more data, this analysis could identify if there are specific analysis areas where the TWG consistently diverges from the regression-based outcome, and if so, are those analysis areas deemed to be in greater or lesser need than a model would predict.

Finally, having shown that the consensus process matters more when the input data are noisy (§4.1.8, Figure 22), we investigate possible drivers of noise in a single case. In Afghanistan, we consider what factors are associated with greater or lower deviation between the consensus outcome and the OLS prediction, and what factors are associated with the consensus outcome generating higher versus lower estimates than the model prediction (§4.1.8, Figure 24). For this single country example, we find the type of ‘noise’ influences the degree of divergence between the consensus and OLS predicted outcomes. We find some evidence that less noisy input FSI data are more likely to generate consensus outcomes that are close to the OLS predictions. This result does not hold for all measures of ‘noise’ in the data, however. For example, we also observe that the greater the range of predictions from the FSI data, the less divergence between the consensus and OLS predicted outcomes. These final results should be taken with some caution given that they are based on findings from a single country and are intended to demonstrate what could be possible with time-series data.

6.2 AFI projections

For AFI projections, we answer the following:

1. What is the relationship between projections and current status classifications?

We find that AFI projections are not consistently higher or lower than TWG current status analyses for that same time period. We find that 4 in 10 projections are consistent with the resulting consensus outcome conducted in the next period. Under and overestimates are about equally split at around 3 in 10 each. Thus, while we find that projections often do not match realized current status analyses, we do not observe systematic over or underestimates of the projections.

Our ability to infer from the projection findings is limited. While our findings do not find systematic evidence of a directional bias (i.e., over- or under-projections relative to current status), we cannot determine why projections differ from current status analyses given the available data. If projections "crowd-in" funding, one may have expected to see current status analyses lower as a result of needs being met. However, limited evidence suggests that funding is not necessarily responsive to declarations of severe food insecurity. Maxwell et al. (2023), in a review of six countries facing severe food insecurity, warn “the relationship between [famine] declarations or warnings and funding is unclear.” Further, when funding does respond to projections, it may keep a situation from deteriorating rather than improving. Thus, future work that carefully collects the timing and amount of funding as well as other shocks could help determine what might cause differences between projections and current status analyses for the same time period. In the recommendations, we propose systematizing a comparison of projections with the realized current status TWG meetings.
6.3 AMN current status and projections

Overall, we find that compared to AFI, data for AMN is much more limited, despite having heard fewer concerns about AMN during our qualitative interviews. This likely reflects the limited understanding that projections and current status may rely on non-representative or less timely data.

Due to limited AMN data on current analyses and projections, we focus on two countries, Kenya and South Sudan, to illustrate possible analyses. We first evaluate the reliability of the evidence used to make classifications. We then evaluate the consistency between representative and non-representative survey-based outcomes.

Second, we evaluate the consistency between the Weight-for-Height Z-score (WHZ) implied classifications and the observed consensus classification. Third, similar to our AFI analyses, we demonstrate an approach to compare the consistency between prior projections and actual current status but do not have temporal overlap in our data.

In our small sample, we find that the majority of the AMN input data are categorized as having less than “High” (3 asterisks) level of reliability. A large number of the observations use historical AMN data (collected within the past five years) for their analyses, and these analyses are more likely to have consensus outcomes that diverge from the reference table. We also find that consensus-based classifications are consistent with the WHZ outcomes implied by the reference table. Without temporal overlap between the prior projection window and the actual current status, we cannot compare AMN projections to outcomes for the same period that they predict and our findings serve as an example of what could be done.

Given these AMN results are based on only two countries for which sufficient data was available, they should be understood as case studies that suggest possible routes for future analyses; these results are not generalizable.

6.4 Limitations

We faced several limitations in our analysis.

First, all accuracy studies of the IPC must contend with a lack of an observable truth, meaning we cannot assess “accuracy” in the purest sense. Rather, we use the Reference Table and underlying food security indicators to examine TWG fidelity to process. With our methods, we can observe and identify differences between our estimated outcomes and the outcomes of the consensus process.

However, we cannot determine the drivers of the observed differences. Data limitations preclude us from accounting for contextualized, contributing factors in our analysis. These factors could (and should) shape TWG outcomes.

Therefore, we underscore: observed differences do not necessarily mean inaccurate or wrong consensus-based outcomes. The differences we document instead suggest that the process generates results that vary with respect to how closely aligned they are with specific FSI input data. Future work that systematically incorporates data on within-TWG factors could more comprehensively model the TWG process, allowing us to identify the influence of FSIs and the convergence process on the convergence outcomes with more precision.

Second, our findings take the available data and the IPC V 3.1 Technical Manual Reference Table as given. The IPC has stewardship over the consensus process but not over the data inputs. In practice, this means that while we had access to aggregated data, we did not observe the underlying analyses or data collection. Thus, any errors in data inputs carry over into IPC analysis, and then into our own analysis contained here. Therefore, just as TWGs can produce incorrect or inconsistent results based on data produced externally to the IPC process, we can as well. Identifying where and when the consensus process diverges from the Reference Table can hopefully suggest places where improvements to the process can be made.

Third, our analysis is constrained by data availability by country. We had access to data from 15 countries and 28 working groups. The IPC has been instituted in 35 countries. Given that we only observe multiple rounds for three countries, our analysis of the drivers of divergence of consensus outcomes is not likely to be generalizable beyond the analysis set. That is, our findings may not apply to countries outside of our sample. With more countries and more time periods within countries, our findings will have more nuance and external validity. Future work that includes more time series within countries could shed light on several analyses (described below in Recommendations). Similarly, access to underlying household level data would allow us to answer questions related to surprising relationships in FSI, for example in Afghanistan.
Fourth, our understanding of data reliability was based on IPC's reliability scores. Other factors can contribute to poor aggregated data (e.g., data collection is outside of IPC's remit. TWGs may only have access to already tabulated data for some indicators and may not be able to verify the analyses, whether the sampling protocols were followed, etc.). We do not observe these possible issues.

Fifth, our analysis did not consider contributing factors, which are important pieces of evidence for TWG analysis. Future research on the role of contributing factors in arriving at consensus outcomes is needed.

Sixth, we do not have many classifications in IPC 1 or any in IPC 5 (see §3.2 Descriptive Statistics). Therefore, we caution that our findings should not be applied to IPC 1 or 5 classifications.

Finally, our methodological approaches to understanding accuracy are some of many possible options. Other approaches may provide valuable insights on different aspects of accuracy. We considered alternatives but found them less appealing than examining fidelity to process.

For example, one alternative would be to compare FEWS NET and CARI to the IPC. Given that the approaches to both FEWS NET analyses and CARI are different both in nature of the process by which they come to conclusions and by their treatment of HFA data, we would not be able to interpret findings from such a comparison. In other words, we would not be able to determine, if FEWS NET analyses and IPC differ, if differences are because of differences in process, different treatment of HFA, or inaccuracy on the part of one agency or the other (or both)? Without being able to rule out HFA or process differences, we would not be able to ultimately say much about accuracy (we do consider HFA in the case of Afghanistan, in §4.1.8 and Annex §3.5). A second approach could be to compare current status to projections for that same period. This approach could interrogate whether the projections match the actual outcomes. However, as discussed earlier in the document, we would hope donors, states, and others respond to projections, leading to lower current status. Incorporating information on HFA could help contextualize gaps between projections and current status outcomes. However, approach would still face limitations inasmuch as being able to accurately forecast all future drivers of food insecurity is not possible.
7. RECOMMENDATIONS

The IPC is uniquely important. The IPC is the gold standard for food security assessment and heavily used for allocations of humanitarian funding. It aims to ensure comparability across space and time and provides a process by which analysts can converge evidence from a variety of information sources. Our recommendations aim to support the important work that TWGs undertake and to assist stakeholders who use IPC findings for critical resource allocation decisions. We recommend that the IPC pair analyses with documentation to support an improved understanding of where TWG classifications deviate from what might be expected based on available data. We identify several possible approaches and propose that the IPC GSU work with TWG members to determine which analysis and documentation best support greater comparability and are not an undue burden.

Below, we first discuss short term recommendations and then long-term recommendations. Most recommendations generally cut across AFI and AMN, although we note where a recommendation reflects either AMN or AFI.

7.1 Short-term recommendations: Ways to support accuracy and perception of accuracy

IPC to systematically study and review thresholds in AFI Reference Table. Given the dispersion of phase outcomes suggested by the recommended FSI thresholds in a single location and time in the vast majority of cases, there may be merit to investigating the thresholds associated with the Reference Table.

IPC to work with TWGs to incorporate automated analyses to support TWGs and IPC’s stakeholders: As the IPC moves towards using the Analysis Platform, an internet platform which will be the standard platform for all analyses, there is an opportunity to use the AP to incorporate automated analyses in support of TWG efforts. We emphasize that such analyses should complement and not replace the consensus process. Human-led consensus processes remain critically important, aiding both buy-in and coordination and reflecting existing limitations in algorithmic approaches and data availability. We recommend that the IPC work with TWGs and its stakeholders to determine which of the following ought to be field tested and or adopted:

- **Projections**: at the end of a TWG analysis, an automated report comparing prior projections to the current status could identify analysis areas with discrepancies or analysis areas where outcomes diverged from assumptions. This could be used to quickly determine where greater documentation or transparency could be beneficial. For example, the TWG could draft a short narrative identifying possible reasons and evidence. Or, if major unanticipated shocks occurred but the prior projection and current status match, the TWG may be able to describe why. This would support learning about the projection process as well as inform stakeholders who used projections to inform decision-making.

- **FSI descriptive information for AFI current status analyses**: IPC could work with its data collection partners to automate descriptive statistics using outcome indicators ahead of TWG analyses. This could include computing current and historical correlations within a country across FSI (3.2.1) or computing the discordance or concordance among classifications implied by the indicators (see 3.2.1). If current correlations between FSI within a country deviate from the observed historical relationships, the TWGs may want to assess and document what could be driving the difference (e.g., timing of data collection, data quality issues beyond reliability scores, etc.) and how the TWG treat the FSI for the current classification. Similarly, two indicators that are highly correlated bring less information to a triangulation exercise than two indicators that are not correlated, and the TWG may not want to equally weigh them in the analysis (all else being equal). Such information could alert the TWG chair as to areas that may warrant more time and more documentation, particularly as the body of historical evidence on FSIs grows within a country.

- **Estimates of weighted predictions for AFI current status analyses**: Ahead of TWG analyses, the IPC could automate a computation of the weighted predictions based on prior TWG information. This could be a starting point for discussion, with the understanding that some FSIs may have data collection concerns for that specific TWG analysis. This could also provide an opportunity for the TWG to document why they deviate from a weighted average for that country. For example, if TWGs down-weight specific FSI (e.g., due to collection issues) in a specific location, that could be documented (i.e., by examining residuals for each analysis area and reporting rationale for high residuals). Alternatively, if this analysis area warrants further discussion with TWGs because FSI should be interpreted to reflect spatial variation, documenting that conversation could be important to ensure that such knowledge is appropriately applied over space and time.
We have seen potential in the area of using machine learning to predict first level outcomes or missing inputs but have not yet seen enough information to know if TWGs will benefit from such approaches. We therefore recommend that the IPC research the following analytical innovations:

- **Predicting missing inputs into the process** to support TWGs (Lentz et al. 2019). For example, in some contexts, when data on contributing factor information (e.g., price information) is missing, predicting outcomes using a machine learning model may provide valuable information. For this effort to be useful, the predictions must not rely on other data already used by the TWG (e.g., by incorporating remote sensing information and would require that modelers understand the context adequately (e.g., predicting prices in a conflict zone without adequate data on conflict would result in inaccurate predictions).

- **Predicting relationships among FSIs, or drivers of FSIs and AMN** prior to TWG analysis these relationships to the TWG before the analysis begins. As data availability expands, future research on modeling AFI and AMN outcome indicators using household level data on food security, SMART data, spatial information, pricing etc. could help TWGs understand influential drivers of food insecurity and acute malnutrition in local contexts. Identifying these relationships might aid TWGs in their AFI and AMN analyses and projections.

**IPC GSU to work with TWGs to identify documentation to support transparency and comparability:** We recognize that documentation is an additional burden on TWGs, which already face packed agendas. However, limited documentation leaves IPC vulnerable to the critique that the consensus process is not transparent and that data are treated inconsistently across TWG analyses. Documentation of certain decisions could build further trust in the consensus process. For example, documentation could discourage TWG members who are “loudest voices in the room” to push for particular preconceived outcomes without fully considering the evidence. The following are a series of approaches that could aid comparability:

- **For AFI current status analyses:** If a TWG determines that certain FSIs should be down-weighted (or upweighted) due to local context, the TWG should document the rationale for this choice. In other words, if LCS is not appropriate in a specific country or IPC analysis area, the reason why should be noted. Over time, harmonizing the contributions of FSIs could aid in building greater within-country consistency in the use of FSI and discourage deployment of particular FSIs in an ad hoc manner.

- **For AMN current status analyses:** GAM prevalence is sometimes drawn from historical data. TWG should document how they use historical GAM data to arrive at current estimates (i.e., describe the assumptions used by the TWG to update GAM prevalence. For example, if they are lowered or raised GAM compared to the historical values, explain what sources and information informed that decision).

- **For AMN contributing factors:** Consistently document sources for specified key drivers and underlying assumptions (i.e., in Figure 129 in Technical Manual V3.1). Document which of the underlying assumptions inform the TWG findings.

- **For AFI and AMN projections:** Projection assumptions in the narrative explanation need better documentation. Currently, IPC projections include key assumptions for the projected period, but IPC reports, after the projection period, do not reflect on which assumptions were correct or moving in the expected direction. As discussed above, compare the prior projected phase and population classifications against the realized current status classifications (e.g., at the end of the TWG analysis). Institute a process to review and document whether assumptions were correct, whether they impacted food security as expected (inasmuch as possible), and whether the situation changed in unanticipated ways. Over time, documenting which assumptions were not met could create a database that would allow for more analysis on which drivers are hardest to anticipate (e.g., if IPC learned that macroeconomic shocks are difficult to predict and their impacts on food insecurity are unclear, macroeconomic shocks could act as a trigger for a new current status analysis). Documentation about which assumptions were or were not met would also enable a case study approach that could help inform learning about projecting food security, particularly in complex cases with multiple factors. We emphasize that forecasting is extremely difficult, and the goal of documentation should be to support learning by the TWG and to support stakeholders who rely on projections for decision-making.

- **Evidence levels and reliability scores:** Aim to resolve discrepancies between evidence level and reliability score. Document remaining discrepancies.

- **Missingness and data reliability:** Document missingness of information and document any data quality concerns that fall outside of the evidence level and reliability score. Over time, commonly occurring issues could be incorporated into the evidence level and reliability score.
IPC to broaden data reliability assessment to consider data quality information in addition to the sample and timeliness of the data collection. Reliability scores are generally assigned at a survey level. Given that food security indicators are generally collected in the same survey, there is lack of variation in these indicators, making the reliability assessments less useful for TWGs analyzing food security indicators with an area. For example, any information about measurement error associated with an indicator could be incorporated into the reliability assessment. We also believe that as TWGs document data concerns, other possible data quality assessment information could emerge.

IPC to invest in and improve AMN data collation processes: Current AMN data organization and reporting processes are a major barrier to analysis. Moving away from excel-based data entry and production of information only as figures rather than also as values in text would help to support the collection of data into a database. The IPC GSU is already undertaking development of a database with automated features and should continue to ensure it meets the needs of the AMN TWG and AMN community.

IPC to communicate on key topics: Although beyond the remit of the accuracy study, we include this here for the sake of completeness. Further described in the Scoping Study, we found some respondents raised concerns that suggested a lack of clarity about how the IPC process worked. IPC GSU produces briefs and FAQs to help people understand the nuances. Additional topics could include (1) an IPC brief on how population estimates are created and, relatedly (2) a description of how the IPC process and the FEWS NET IPC compatible process differ regarding population estimates and treatment of HFA.

Finally, we recommend to the broader IPC community: Invest in data collection and data sharing. While beyond the remit of the IPC, the quality of the data underpins the consensus process. Investing in data collection and analysis and making raw data available to TWGs when needed will help improve the accuracy of the IPC. Two specific data-sparse areas include the lack of AMN data and the lack of consistent HFA data. Humanitarians, donors, and state actors should commit to documenting and sharing HFA information in a timely fashion. Given the data limitations faced by AMNTWGs (e.g., lack of timely, representative data), making raw SMART data available, which includes information beyond GAM, could help AMN TWGs seeking to rely on contributing factor evidence.

7.2 Long-term recommendations: Undertake an accuracy study in five years

Our final set of recommendations is that the IPC GSU consider undertaking a similar accuracy study in roughly five years. We recommend that the future study updates the current AFI analyses as well as undertake a more complete analysis of AMN. As discussed above, a significant amount of project time was spent collecting data and cleaning it. If the GSU wants to maintain data for internal management and for research, continuing to invest in data management and developing data sharing protocols will be both necessary and beneficial.

Some research questions emerged from the Scoping Study that were not possible with the current data available. To aid a future study’s analyses, we recommend that IPC continues to collect or to consider implementing processes now to collect data unavailable for this study.

First, time series data within countries that are longer than are currently available will help examine several related issues. A concern raised during the qualitative Scoping Study was the degree to which population numbers are a function of prior classification for both AMN and AFI. Respondents explained that it is difficult for TWGs to start with a zero-base population but relying on prior population estimates introduces a risk of carryover of inflated figures or difficulties in dropping figures down.

Relatedly, longer time series would enable investigations into whether certain analysis areas have consistently higher (or lower) classifications and populations than would be expected, which could suggest some external influences within a country. Some respondents reported an interest in determining the value of each additional FSI variable. Longer time series within TWGs could help show how FSIs are used within local contexts and whether they are used in a consistent manner, which when paired with an analysis of the weights, could indicate the marginal information provided by each FSI. Finally, longer time series within countries could enable comparisons on whether TWGs within a country treat FSIs consistently. We might expect variations over space due to the contextual nature of food security indicators. But should we expect these variations within a country over time?
Second, some respondents reported expecting IPC AFI and FEWS NET’s IPC compatible classifications to be consistent. We could not answer with available data; the scope to do so depends on multiple data challenges. To examine the degree of consistency requires spatially consistent analysis areas across IPC and FEWS NET analyses. There is imperfect overlap. Additional time and resources would be needed to match these analysis areas and to investigate how to treat partially overlapping analysis areas. Relatedly, respondents suggested that differences in frequency of analyses (and therefore the data sources used) and or how IPC and FEWS treat humanitarian food assistance (HFA) could contribute to differences in population estimations and classification. Efforts to collect and analyze how HFA data enter into each analysis could help identify patterns in population estimates or differences in phases (mostly at extreme food insecurity levels). Similarly, comparisons between projections and current status analyses will be more robust if analyses can account for HFA allocations occurring in response to a projection. Such data are necessary to address concerns the specific concerns raised by stakeholders in this study. However, we recognize that public release of HFA data remains challenging; collection of HFA data would require significant support from Food Security Cluster members and for that reason may not be feasible.

Third, AMN data collection and documentation will help create a more robust picture of how data are used by TWGs and what could help support them in their work.

Additional qualitative data could also be valuable to understand how perceptions about the IPC’s accuracy have changed and what new concerns have emerged. For example, our Scoping Study indicated respondents were concerned that population estimates were high or that influential stakeholders within or across TWG analyses overclassified certain areas to draw in support. We did not find evidence of either of these, but due to our relatively small sample of countries and short time series duration, we cannot conclusively exclude this explanation. Over time, IPC may want to identify if these or other perceptions persist or if other concerns have emerged; this study is an important step in examining whether these perceptions are accurate. Regular evaluations create an opportunity for ongoing learning about IPC outcomes and stakeholders’ understandings of them, and for identifying opportunities to support food security analysts undertaking IPC analyses and IPC’s end users.
ANNEX 1: What AFI data were we able to include and why?

1.1 AFI outcomes from the IPC-API

We retrieved ‘published’ AFI outcomes from the IPC-API. This includes information such as the total population analyzed, the number and percentage of people falling into each phase outcome (ranging from 1 to 5) by country, analysis period, and area. These figures are available for the current period, as well as the first and (in some cases) second projections (projecting one and two quarters ahead) as long as the analysis is performed. The data includes the IPC phase value (ranging from 1 to 5, as specified by the IPC Technical Manual V3.1) based on the population share. We further obtained geospatial information from the API to visually present our analysis results on a map in §4.1.7.

Our use of AFI outcomes involves two distinct approaches. First, we examine the convergence process by focusing on a subset of the sample that includes data on underlying food security indicators (FSI) obtained separately from the Consolidated Data Tools (CDT). Second, for analyses that primarily rely on IPC phase outcomes and population estimates, such as identifying bunching (§4.1.2) and comparing projections (completed in time t-1) against realized current status outcome comparison (completed in time t) (§ 4.2), we extract specific cases from the entire set of classifications accessible through the IPC-API. Detailed information on the sampling procedures and outcomes for both approaches can be found in the next two sections.

1.2. Sampling Procedure 1: Evaluating the Convergence Process

We collect the underlying FSI using the Consolidated Data Tools (CDT), which is the tool where IPC gathers evidence for published outcomes. The CDT provides information on the population share corresponding to each phase outcome based on key FSI such as Food Consumption Score (FCS), Household Dietary Diversity Score (HDDS), Household Hunger Scale (HHS), reduced Coping Strategies Index (rCSI), and Livelihood Coping Strategies (LCS). Each FSI uses its own threshold defined by the IPC Technical Manual V3.1 (see IPC Technical Manual V3.1 - Protocol 2.2.). We merge AFI outcomes with underlying FSI based outcomes from the CDT based on analysis date and area name.

Table A1. Example of evidence (e.g. FSI based population share (%)): screenshot from the CDT

The table illustrates how the population distribution by FSI and phase is collected in the CDT. It provides information on the share of the population corresponding to each FSI category.
Table A2. Summary of available FSI for country x round level analyses

<table>
<thead>
<tr>
<th>Analysis</th>
<th>L_FCS</th>
<th>FCS</th>
<th>rCSI</th>
<th>HHS</th>
<th>LCS</th>
<th>HDDS</th>
<th>FIES</th>
<th>Sample</th>
<th>Available FSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan, 2021-03</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>45</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Afghanistan, 2021-09</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>45</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Afghanistan, 2022-03</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>45</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Afghanistan, 2022-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>45</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>CAR, 2021-04</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>65</td>
<td>FCS/RCSI/HHS/LCS</td>
</tr>
<tr>
<td>CAR, 2021-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>62</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>CAR, 2022-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>64</td>
<td>FCS/RCSI/HHS/LCS</td>
</tr>
<tr>
<td>DR Congo, 2021-02</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>166</td>
<td>FCS/RCSI/LCS/HDDS</td>
</tr>
<tr>
<td>DR Congo, 2021-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>179</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>DR Congo, 2022-08</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>184</td>
<td>FCS/RCSI/HHS/LCS/</td>
</tr>
<tr>
<td>Djibouti, 2022-03</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>15</td>
<td>FCS/RCSI/LCS/HDDS</td>
</tr>
<tr>
<td>Ethiopia, 2021-05</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>FCS/RCSI/HHS/LCS/</td>
</tr>
<tr>
<td>Guatemala, 2022-03</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>NA</td>
<td>22</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>Haiti, 2021-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>31</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Haiti, 2022-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>32</td>
<td>FCS/RCSI/HHS/HDDS</td>
</tr>
<tr>
<td>Kenya, 2021-02</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>NA</td>
<td>23</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>Kenya, 2021-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>NA</td>
<td>23</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>Lebanon, 2022-09</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>NA</td>
<td>52</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>Madagascar, 2021-05</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>10</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Madagascar, 2022-04</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>16</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Madagascar, 2022-11</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>21</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Mozambique, 2021-11</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>63</td>
<td>FCS/RCSI/LCS/HDDS</td>
</tr>
<tr>
<td>Pakistan, 2021-03</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>NA</td>
<td>19</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>Pakistan, 2021-10</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>25</td>
<td>FCS/RCSI/LCS/HDDS</td>
</tr>
<tr>
<td>South Sudan, 2022-10</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>77</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Sudan, 2021-04</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>NA</td>
<td>179</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>Sudan, 2022-05</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>o</td>
<td>NA</td>
<td>NA</td>
<td>179</td>
<td>FCS/RCSI/LCS</td>
</tr>
<tr>
<td>Yemen, 2020-10</td>
<td>NA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>NA</td>
<td>181</td>
<td>FCS/RCSI/HHS/LCS/HDDS</td>
</tr>
<tr>
<td>Total TWGs</td>
<td>3</td>
<td>28</td>
<td>28</td>
<td>17</td>
<td>27</td>
<td>16</td>
<td>2</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Total Observations</td>
<td>135</td>
<td>1855</td>
<td>1844</td>
<td>1104</td>
<td>1823</td>
<td>930</td>
<td>56</td>
<td>1881</td>
<td></td>
</tr>
</tbody>
</table>

The table presents a summary of the sample used for evaluation, providing details on the size and availability of FSI for each analysis. The sample encompasses 1881 IPC livelihood analysis area classifications from 15 different countries, involving a total of 28 TWG analyses. The initial sample size was 1901; however, 20 cases were excluded from the analysis as they fell outside the 1 to 5 range, indicating either non-analyzed or insufficient evidence. NA indicates not available; ‘o’ indicates available.

FSI availability was examined by considering whether an indicator was available for 50% or more in the specific analysis. For instance, HDDS-based outcomes were available for 25 out of 45 districts in Afghanistan (2022-09) and thus labeled as ‘o’ (available). On the other hand, HHS-based outcomes were only available for 48 out of 166 samples in the Democratic Republic of Congo (2021-02), resulting in a ‘NA’ label (not available). We do not use FIES and lower bound FCS in this analysis because there were very few cases of them in our sample.
1.3. Sampling Procedure 2: Identifying Distributional (Bunching) Patterns and Projection (t-1) and Realized Current Status (t) Outcome Comparison

We select specific cases from the complete set of classifications available through the IPC-API for analyses that focus on IPC phase outcomes and population estimates. This includes identifying bunching, and comparing projections (t-1), against realized current status outcomes (t). The initial set of classifications accessible through the IPC-API consists of 9,461 samples. However, we exclude samples that fall outside the 1 to 5 range, indicating either non-analyzed or insufficient evidence, resulting in a reduced sample size of 8,464.

From the reduced sample, we further filter for cases that have both current status and projection data, resulting in a sample size of 7,360 classifications encompassing 32 countries and 171 TWG analyses. This subset is used for identifying bunching patterns (§4.1.6). Then, we subset the cases where the current status outcomes fall within the validity period of the first projection outcomes, resulting in a sample size of 3,068 involving 19 countries and 55 TWG analyses. This subset is used to examine the accuracy of IPC’s projections (§4.2). Finally, the sample that includes the food security indicators is 1,881, drawn from 15 countries and 28 TWG analyses.

Table A3. Summary of sampling procedures

<table>
<thead>
<tr>
<th>Use</th>
<th># Number of Classifications</th>
<th># of countries</th>
<th># of TWG analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Universe of classifications</td>
<td>9,461</td>
<td>30</td>
<td>211</td>
</tr>
<tr>
<td>(2) Classifications classified between 1-5</td>
<td>8,464</td>
<td>30</td>
<td>190</td>
</tr>
<tr>
<td>(3) Final Sample with both current and projection status outcomes are available</td>
<td>7,360</td>
<td>30</td>
<td>171</td>
</tr>
<tr>
<td>(4) Final Sample with current status outcomes that match the validity period of the first projection outcomes</td>
<td>3,068</td>
<td>19</td>
<td>55</td>
</tr>
<tr>
<td>(5) Final Sample of classifications with FSIs drawn from CDT</td>
<td>1,881</td>
<td>15</td>
<td>28</td>
</tr>
</tbody>
</table>

The table provides a summary of the sampling procedures, including the sample size and the use of each sample.

1.4. Reliability Score for AFI

Each Consolidated Data Tools (CDT) analysis includes reliability scores (R) at the area or district level. These scores are determined based on two components: the soundness of the method (M) and the time relevance (T) of the data, each ranging from 1 (limited) to 2 (good) followed by the alphabet identifier M and T. The reliability score in the analysis ranges from R1- to R2, with R2 indicating the highest level of reliability and R1- representing the lowest level of reliability (please refer to IPC Technical Manual V3.1 - Protocol 2.4. for more detailed information). While the reliability score can vary within a TWG or analysis, it does not vary across food security indicators within a district, as the available FSI appear to be usually collected in a single survey for each region.

Table A4. Example of reliability score table: screenshot from CDT

The table illustrates how the reliability score and its sub-components, such as time relevance and soundness of method, are collected in CDT for each sampling unit.
For a comprehensive description of the variation in reliability scores, including the sub-components of the score (timeliness and soundness of method) by TWG analysis, refer to Table A5.

**Table A5. Variation of reliability score by country x round**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Reliability Score</th>
<th>Timeliness</th>
<th>Soundness of Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consistency</td>
<td>Unique Cases</td>
<td>Consistency</td>
</tr>
<tr>
<td>Afghanistan, 2021-03</td>
<td>Consistently High</td>
<td>R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Afghanistan, 2021-09</td>
<td>Consistently High</td>
<td>R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Afghanistan, 2022-03</td>
<td>Consistently High</td>
<td>R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Afghanistan, 2022-09</td>
<td>Consistently High</td>
<td>R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>CAR, 2021-04</td>
<td>Consistently Low</td>
<td>R1-</td>
<td>Consistently Low</td>
</tr>
<tr>
<td>CAR, 2021-09</td>
<td>Not Consistent</td>
<td>R2, R1+ Not Available</td>
<td>Consistently High</td>
</tr>
<tr>
<td>CAR, 2022-09</td>
<td>Not Consistent</td>
<td>R1+ R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>DR Congo, 2021-02</td>
<td>Not Available</td>
<td>Not Available</td>
<td>Not Available</td>
</tr>
<tr>
<td>DR Congo, 2021-09</td>
<td>Not Consistent</td>
<td>R1+ R2, Not Available</td>
<td>Consistently High</td>
</tr>
<tr>
<td>DR Congo, 2022-08</td>
<td>Not Consistent</td>
<td>R1+ R2, Not Available</td>
<td>Not Consistent</td>
</tr>
<tr>
<td>Djibouti, 2022-03</td>
<td>Not Consistent</td>
<td>R1+ R2, R1- Not Available</td>
<td>Not Consistent</td>
</tr>
<tr>
<td>Ethiopia, 2021-05</td>
<td>Consistently Low</td>
<td>R1-</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Guatemala, 2022-03</td>
<td>Not Consistent</td>
<td>R2, R1+</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Haiti, 2021-09</td>
<td>Consistently High</td>
<td>R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Haïti, 2022-09</td>
<td>Consistently High</td>
<td>R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Kenya, 2021-02</td>
<td>Not Consistent</td>
<td>R2, R1+</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Kenya, 2021-09</td>
<td>Not Consistent</td>
<td>R2, R1+</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Lebanon, 2022-09</td>
<td>Consistently Low</td>
<td>R1+</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Madagascar, 2021-05</td>
<td>Consistently Low</td>
<td>R1+</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Madagascar, 2022-04</td>
<td>Not Consistent</td>
<td>R2, R1+ Not Available</td>
<td>Not Consistent</td>
</tr>
<tr>
<td>Madagascar, 2022-11</td>
<td>Consistently High</td>
<td>R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Mozambique, 2021-11</td>
<td>Not Consistent</td>
<td>R1+ Not Available</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Pakistan, 2021-03</td>
<td>Consistently Low</td>
<td>R1-</td>
<td>Consistently Low</td>
</tr>
<tr>
<td>Pakistan, 2021-10</td>
<td>Not Consistent</td>
<td>R1+ R2</td>
<td>Consistently High</td>
</tr>
<tr>
<td>South Sudan, 2022-10</td>
<td>Consistently Low</td>
<td>R1-</td>
<td>Consistently Low</td>
</tr>
<tr>
<td>Sudan, 2021-04</td>
<td>Not Consistent</td>
<td>R1+ Not Available</td>
<td>Not Consistent</td>
</tr>
<tr>
<td>Sudan, 2022-05</td>
<td>Consistently Low</td>
<td>R1+</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Yemen, 2020-10</td>
<td>Not Consistent</td>
<td>R1+ R1- Not Available</td>
<td>Not Consistent</td>
</tr>
</tbody>
</table>

This table provides a comprehensive description of the variation in reliability scores, including the sub-components of the score (timeliness and soundness of method) by TWG analysis.
ANNEX 2: What AMN data were we able to include and why?

2.1. Data Collection and Limitation

Our collation of AMN data was limited, reflecting substantial data cleaning and data trustworthiness challenges. In the following section, we provide detailed documentation of the limitations encountered during the data collection process.

Data Source 1: IPC reports

Initially, we sought to obtain data from IPC AMN reports. However, this approach presented significant challenges for several reasons. First, key information such as population estimates of global acute malnutrition (GAM) (%), underlying indicators like MUAC or WHZ, and evidence levels were often missing for many analysis areas and time periods. Second, the data were primarily available in PDF or JPG format, making it difficult to extract them into a machine-readable format. Translating the data from PDFs is a laborious process as the figures within the PDFs had varied structures and were written in different languages and would require manual entry.

Screenshots from the Central African Republic (October 2022) AMN report (see Figure A1) highlight the difficulties in capturing AMN data (IPC CAR 2022). The AMN outcomes, specifically the classification phases, were only available as images. Additionally, underlying indicators such as GAM were often presented in tables spread across multiple pages in the PDFs, requiring manual entry. The evidence levels of AMN data, indicated by asterisks on the map, needed to be manually converted into usable data, further contributing to the labor-intensive nature of the process.

Figure A1. Screenshots from AMN report (Central African Republic - October 2022)
Our observations indicate that these challenges in compiling AMN data are common across various reports, including cases missing crucial pieces of information. In summary, merging all the necessary information at the IPC analysis area level, including AMN outcomes, population percentages, GAM percentages, and evidence levels, proved to be challenging due to the diverse formats in which the data were presented in the PDFs. Therefore, we sought alternative sources of AMN data.
Data Source 2: AMN worksheets

Based on data limitations, we revised our objective for the AMN analysis to conducting three proof-of-concept analyses in Kenya and South Sudan, using excel-based worksheets generated to support AMN TWGs. The excel worksheets have separate sheets for each step of the protocol. However, TWGs slightly change the format of the spreadsheet, making it challenging to extract data via automated processes. For example, extracted values from cell ‘A2’ in ‘STEP 4’ across worksheets provided inconsistent information. Thus, we worked with two countries and manually combined their data. In Kenya, we have data from three rounds of AMN classifications (2020, 2021, 2022), providing a substantial number of classifications and utilizing different data sources. In South Sudan, we have data from two time periods within a single year (2022), allowing us to examine both current and projected classifications.

Problems exist also in the AMN worksheets. We identified two main issues in STEP 3A of the analysis. First, there is an inconsistency between the measures used to capture data reliability, namely the evidence level (see Figure 132 in IPC Technical Manual V3.1) and reliability score (see Figure 131 in IPC Technical Manual V3.1). The evidence level is denoted by asterisks and is a summary of the components of the evidence reliability, which is denoted by R1, R2, etc. In the table below, we see that 8 cases from the 56 Kenyan samples have an reliability score of R2, indicating good time relevance and good soundness of methods. R2 is reserved for GAM evidence based entirely on WHZ scores from representative surveys and reflecting current conditions. However, the evidence level for four of those cases is denoted as having either one or two stars. Two stars (indicated as ** evidence level) denote cases relying on two pieces of contributing factor evidence and either R1+ GAM using WHZ or MUAC data that is not representative. One star (indicated as * evidence level) denotes cases relying on two pieces of contributing factor evidence and R1 - GAM using less timely WHZ score evidence. It is unclear why R2 data would receive an evidence level other than *** or why R1+ would receive an evidence level other than ** based on the Technical Manual V3.1 information. Given this inconsistency between the evidence reliability and evidence level, we rely on the evidence level published by IPC in their reports.

Table A6. Inconsistency between evidence level and evidence reliability for Kenya AMN data

<table>
<thead>
<tr>
<th>Evidence Reliability</th>
<th>Evidence Level</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1+</td>
<td>**</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>5</td>
</tr>
<tr>
<td>R2</td>
<td>***</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>2</td>
</tr>
</tbody>
</table>

Second, there are inconsistencies between the Technical Remarks and the Prevalence (%) of GAM based on WHZ. In some cases, the Technical Remarks suggest a phase outcome of Phase 1 based on historical data, but the actual outcome listed is Phase 2 without sufficient justification. Similarly, there are instances where the Technical Remarks mention the use of historical data, but no historical data is available. Moreover, there are discrepancies between the Technical Remarks and Step 4 regarding the reliance on WHZ and or MUAC. To address these issues, we conduct a cross-check of the information with the IPC report and other relevant pieces of information in the worksheet to determine the most accurate and consistent information to incorporate into our analysis.

Due to the limited sample size and coverage, we minimize discussion on the data reliability in our analysis and refer to the guidance provided in the IPC Technical Manual 3.1, where these issues are extensively addressed.

2.2 Data Sampling

To examine the accuracy of acute malnutrition (AMN), we analyze data extracted from the AMN worksheets, which involves a multi-step decision-making process to determine AMN outcomes. Our evaluation relies on evidence from various sources, including the prevalence of global acute malnutrition (GAM) based on WHZ or MUAC (Step 4), the current status classification with reliability scores and justifications (Step 4), and projection classifications (Step 8). Kenya and South Sudan serve as case studies to demonstrate three analyses to examine accuracy or identify unforeseen inconsistencies, but results should not be extrapolated to the universe of AMN classifications. Table 4 is the summary of the data used in the analysis.

IPC GSU agrees this is an error. R2 classifications should have three stars. R1+ should have only two stars. Automated checks (discussed in recommendations) could flag these in real time.
ANNEX 3: Additional Findings

3.1. FSI vs. Consensus-based 3+ Population (%) - country specific results

Figure A2 below presents results similar to Figure 8 in §4.1.1. by country for Afghanistan and Yemen. Comparing across country-specific figures, we see somewhat different relationships across food security indicators; we discuss this in our regression results in §4.1.5.

The graphs present the 3+ population classified for all TWG analyses with five food security indicators for two countries, Afghanistan (n=160) and Yemen (n = 176). In both, the x-axis represents the share of the population the TWG assesses to be in phase 3+ through the consensus process. The y-axis represents the share of the population who would be assigned to IPC 3+ determined only by the food security indicator data (i.e., the population implied by each food security indicator). The gray dotted 45-degree line depicts the hypothetical situation where the actual outcome from the consensus share in 3+ (%) is equal to the share implied to be 3+ (%) by the food security indicators. The thick black line represents the arithmetic mean in which we weight equally the population implied to be 3+ by each food security indicator. The gray shaded area shows the distribution of the cases classified and should be read against the right-hand y-axis.
3.2. FSI vs. Consensus-based 3+ Population (%) - cases with 4 FSI

Figure A3 presents results similar to Figure 8 for other combinations of food security indicators (e.g., FCS, LCS, rCSI, and HHS). Results are consistent with the sample with all five indicators.

Figure A3. AFI – current status – consensus-based 3+ population (%): 4-FSI Group (without HDDS)

The graph presents the 3+ population classified for all TWG analyses with four food security indicators. The x-axis represents the share of the population the TWG assesses to be in phase 3+ through the consensus process. The y-axis represents the share of the population who would be assigned to IPC 3+ determined only by the food security indicator data (i.e., the population implied by each food security indicator). The gray dotted 45-degree line depicts the hypothetical situation where the actual outcome from the consensus share in 3+ (%) is equal to the share implied to be 3+ (%) by the food security indicators. The thick black line represents the arithmetic mean in which we weight equally the population implied to be 3+ (%) by each food security indicator. The gray shaded area shows the distribution of the cases classified and should be read against the right-hand y-axis. (n = 388)
The graph presents the 4+ population classified for all TWG analyses with five food security indicators \((n = 695)\). The x-axis represents the share of the population the TWG assesses to be in phase 4+ through the consensus process. The y-axis represents the share of the population who would be assigned to IPC 4+ determined only by the food security indicator data (i.e., the population implied by each food security indicator). The gray dotted 45-degree line depicts the hypothetical situation where the actual outcome from the consensus share in 4+ (%) is equal to the share implied to be 4+ (%) by the food security indicators. The thick black line represents the arithmetic mean in which we weight equally the population implied to be 4+ (%) by each food security indicator. The rCSI line indicates that the indicator cannot differentiate between IPC 4 and 5. The gray shaded area shows the distribution of the cases classified and should be read against the right-hand y-axis. The sample consists of the following countries: Afghanistan, Democratic Republic of Congo, Haiti, Madagascar, South Sudan, and Yemen.
3.4. Difference between IPC 3+ population (%) and the arithmetic mean of the FSI implied 3+ population (%) - country/TWG specific results

Figure A5. Distribution of the difference between consensus-based 3+ population (%) and the arithmetic mean of FSI implied 3+ populations (%) by country

The histogram presents the observed differences in the share of the population assessed in IPC phase 3+ and by the arithmetic mean of the FSIs. We show results by country. Blue bars depict analyses by the IPC Country that classified a smaller population as experiencing acute food insecurity (IPC 3+) than the population according to the simple average of the FSI. That is, they assessed fewer people at 3+ than implied by a simple average of the FSI. The green bar depicts IPC population analyses that are consistent with the arithmetic mean. Red bars depict IPC population analyses that are less conservative than the simple average - that is, where the IPC assessed that a larger population was in phase 3+ than implied by the simple average of the FSI. Sample includes all analyses, regardless of the number of FSI available (n=1881).

Figure A6. Distribution of the difference between consensus-based 3+ population (%) and the arithmetic mean of FSI implied 3+ populations (%) by country x round

The histogram presents the observed differences in the share of the population assessed in IPC phase 3+ and by the arithmetic mean of the FSIs. We show results by country x round. Blue bars depict analyses by the IPC TWG that classified a smaller population as experiencing acute food insecurity (IPC 3+) than the population according to the simple average of the FSI. That is, they assessed fewer people at 3+ than implied by a simple average of the FSI. The green bar depicts IPC population analyses that are consistent with the arithmetic mean. Red bars depict IPC population analyses that are less conservative than the simple average - that is, where the IPC assessed that a larger population was in phase 3+ than implied by the simple average of the FSI. Sample includes all analyses, regardless of the number of FSI available. (n=1881)
3.5 Correlation of residuals to underlying characteristics of the input data

In what follows, we consider which underlying characteristics of the FSI data are correlated with the absolute value of the residuals in the findings for four rounds of data in Afghanistan (§4.1.8). The residuals are from a regression of FSI data on consensus-based population outcomes. Through this analysis, we can observe whether residuals have systematic patterns over space and time or appear random, which might be consistent with them merely being noise. More information on the analysis is available in §2.4.4.

In Figure A7, we first observe that the higher the maximum phase predicted by the FSIs, the lower the divergence between the weighted FSI average and the consensus outcome, which suggests that TWGs place relatively higher weights on FSI data that predict high phases. On the other hand, the minimum phase predicted by the FSIs is not correlated with the deviation between the FSI predictions and the consensus outcome.

Next, we explore the range of phase outcomes suggested by the FSIs. For example, we might expect that the greater the range of the phase outcomes predicted by the underlying FSIs for that IPC analysis area at a given time, the greater the absolute value of the deviation between the predicted FSI phase and the phase outcome of the convergence process (i.e., the absolute values of the residuals). This expectation does not appear to hold. As can be observed by the third bar, the difference between the maximum and minimum phase predicted by different FSIs for that location is negatively correlated with the deviation in convergence and FSI predictions. Next, we observe the somewhat counterintuitive result that the higher the HFA, the lower the agreement between the consensus process and the weighted average result from the FSIs. We might anticipate that analysis areas receiving substantial food assistance might be places where the consensus process deviates more from the algorithmic outcome because either the TWG adjusts their analysis because of the HFA or that there are extenuating circumstances that drive both the HFA and the analysis. Instead, we see that these areas receive analyses more closely aligned to the algorithmic results.

We find the greater the percent of FSIs that predict the same outcome, the higher the divergence between the consensus and algorithmic outcome. This result suggests that having many FSIs point to the same outcome does not drive the consensus process to follow suit.

The following set of bars show the correlation between the gap between the consensus-based outcome and each FSI predicted outcome, and the absolute value of the residuals. We see that the greater the gap between the consensus outcome and FCS, HDDS, HHS and LCS, the greater the divergence between the consensus outcome and the predicted value coming from the FSIs. This result is what we would expect. The only counterintuitive result is for rCSI, where the greater the deviation between rCSI and the consensus outcome, the smaller the deviation between the result predicted by the FSIs in total. This result suggests that there is not a great weight placed on rCSI, or that it is often distrusted in the setting of the percent of population in phase 3+.

For 4 of the 5 FSIs (FCS, HDDS, HHS, and LCS), predicting a higher phase is correlated with a greater deviation in the resulting versus the FSI-predicted phase (i.e., larger absolute value of the residuals). This finding is again suggestive that the resulting phase classifications are conservative, in that it is when the FSIs indicate high phases that the percent of population in IPC Phase 3 or above (crisis or worse) from the consensus process diverges more from the percent of population in IPC Phase 3 or above (crisis or worse) suggested by the FSIs. This finding is consistent with the figures of the FSI predictions above, where the HHS predictions tracked the convergence outcomes more closely than most other FSIs.
The y-axis is the Pearson’s correlation between the absolute value of residuals estimating the population in IPC Phase 3 or above (crisis or worse) over all four rounds of data from Afghanistan as an outcome of the consensus process versus the predictions generated from the weighted average of the percent of population in IPC Phase 3 or above (crisis or worse) implied by the FSIs and the variables listed on the x-axis (n = 160).

In Figure A8, we explore how these same factors are correlated with whether the consensus population estimate is higher or lower than that predicted by the underlying FSIs. First, we observe that the maximum phase suggested by the FSIs is correlated with the convergence outcome being lower than the predicted FSI outcome. This result may be largely mechanical since when one FSI suggests a high phase, the model-predicted population (%) is also likely to be high, increasing the probability that the consensus process outcome is lower. Second, the larger the range in FSI implied phase outcomes, the more the consensus population tends to be lower than the model predicted population. This suggests that noisy input data are related to conservative consensus outcomes. Similarly, having a greater consensus of evidence from the FSIs generally results in a higher convergence outcome relative to the weighted average outcome predicted by the FSIs. As can be seen in column 5, the higher the percent of the FSIs that predict the convergence outcome, the more likely the convergence outcome is higher relative to the FSI outcome.

As one might expect, higher rates of food assistance are associated with higher consensus outcomes relative to the weighted average FSI predictions. More food assistance may cause the FSI outcomes to underestimate the true level of food insecurity that would exist without assistance, or it may suggest that funders observe other information that suggest need, both of which would be consistent with a TWG identifying a larger portion of the population in IPC Phase 3 or above (crisis or worse).

When the consensus phase classification is larger than the phase classified by any particular FSI, the consensus phase classification tends to predict a higher level of classification than the weighted average of the FSIs. This result may suggest that analysis areas where the consensus process tends to identify a higher phase than suggested by one FSI, it tends to be higher than the phase predicted by other FSIs as well. Also notable is that the correlations of the residuals with the prediction from each FSI are consistent with the figure A2 above: that FCS tends to estimate a higher percent versus rCSI tends to estimate a lower percent of people in need than the percent resulting from the convergence process. We can observe this result by noting that when FCS implies the consensus outcome, the more likely the consensus outcome is higher than the weighted FSI average. Conversely, when rCSI implies the consensus outcome, the consensus outcome tends to fall below that predicted by the weighted average FSI.

Last, for most FSIs, if they predict a higher outcome, the convergence outcome also tends to be higher, relative to the weighted average FSI prediction. Notably FCS is the outlier, where if FCS predicts a higher outcome, the consensus outcome tends to be lower than the weighted average FSI prediction. This latter outcome makes sense given that FCS tends to predict higher outcomes than both other FSIs and the consensus process.
Figure A8. Pairwise correlation between the residuals and potential sources of noise in Afghanistan

The y-axis is the Pearson’s correlation between the value of residuals estimating the population in IPC Phase 3 or above (crisis or worse) over all four rounds of data from Afghanistan as an outcome of the consensus process versus the predictions generated from the weighted average of the percent of population in IPC Phase 3 or above (crisis or worse) implied by the FSIs and the variables listed on the x-axis (n=160).

3.6. AFI Realized Current Status Analysis Classifications (RCSA) at time (t) vs. Projected Status (PS) from the prior quarter (t-1)

The figure is a country specific comparison of PS and RCSA for the sample TWG classifications with projection window overlapping with realized current status. The Figure presents the percentage (%) of each incidence (PS < RCSA, PS = RCSA, or PS > RCSA) based on relative frequencies from samples where we can match PS t-1 and RCSA t. Green indicates when the Projected Status classification for each case matches the Realized Current Status Analysis for that case. Blue indicates when the Projected Status for each case was below the Realized Current Status Analysis. Red indicates when the Projected Status for each case was above the Realized Current Status Analysis (n = 3068).
The figure is a TWG specific comparison of PS and RCSA for the sample TWG classifications with projection window overlapping with realized current status. The figure presents the percentage (%) of each incidence (PS < RCSA, PS = RCSA, or PS > RCSA) based on relative frequencies from samples where we can match PS t-1 and RCSA t. Green indicates when the Projected Status classification for each case matches the Realized Current Status Analysis for that case. Blue indicates when the Projected Status for each case was below the Realized Current Status Analysis. Red indicates when the Projected Status for each case was above the Realized Current Status Analysis (n = 3068).

3.7. Distributional analysis: Country specific bunching results

Figure A11 suggests that the distribution of the percent population identified in IPC Phase 3 or above (crisis or worse) or phase 3+ population (%), exhibits more ‘bunching’ away from the 20% cut-off than others, where the 20% cutoff is illustrated by the red vertical line. Specifically, we observe potential bunching in Lebanon, Guatemala and the Democratic Republic of Congo.
The figure illustrates the kernel density estimates of the population (%) in IPC Phase 3 or above (crisis or worse) (IPC 3+) identified by the consensus process (illustrated in black). The 20 percent population in IPC Phase 3 or above (crisis or worse) used as a cut-off for phase 3 is illustrated by a vertical red line (n = 7271).
In the case of Lebanon (see Figure A12), where only one IPC-AFI report is available for September 2022, we observe a bunching pattern, characterized by a kink in the distribution just below the 20% cutoff, specifically among the non-refugee group. However, no such pattern is evident among the Refugees. This variation in bunching behavior may be attributed to the cautious approach taken by the TWG in declaring Phase 3 for the Non-refugee group, particularly considering that it was their first experience conducting the analysis. Other factors may have influenced this report and given that the difference is based on a single TWG analysis, it should not be treated as generalizable.

![Figure A12. Distribution of the population (%) in IPC Phase 3 or above (crisis or worse) (IPC 3+) - Lebanon (Refugee VS. Non-Refugee Group)](image)

The figure shows the kernel density estimates of the population (%) in IPC Phase 3 or above (crisis or worse) (IPC 3+) as determined by the consensus process in Lebanon for September 2022. The distribution for the refugee group is represented by black line, while the non-refugee group is shown in gray. A vertical red line indicates the 20 percent population threshold used to determine phase 3 classification (n = 52).

### 3.8 Additional results

![Figure A13: FSI implied 3+ population (%) over time in Afghanistan](image)

This figure illustrates the average and standard deviations of the population in IPC Phase 3 or above (crisis or worse) (%) determined by five FSIs across four rounds of analysis in Afghanistan. The y-axis represents the average and standard deviations of the 3+ population (%). Each panel sequentially showcasing results using FCS, HDDS, HHS, LCS, and rCSI from left to right (n = 180).
This figure demonstrates the difference between the consensus-based phase classification and the FSI-implied phase classification. The categories indicate the size of the difference. Gray = 0 indicates no difference between 'country x round' level phase classification and the FSI implied phase classification. As colors become darker blue, the consensus-based phase classification is lower than the FSI implied phase classification. As colors become darker red, the classification is higher than what is implied by the FSI. The x-axis is country by (x) round. The y-axis indicates share of classifications in each category for each analysis (n=1881).
Table A7: FSI weights from the OLS regression models that estimate consensus-based 3+ population for creating Figure 14 - All-5-FSI group

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Coefficients</th>
<th>Standard errors</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCS 3+ population (%)</td>
<td>0.166</td>
<td>0.261</td>
<td>Madagascar, 2022-04</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.184</td>
<td>0.266</td>
<td>Madagascar, 2022-04</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.278</td>
<td>0.223</td>
<td>Madagascar, 2022-04</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>-0.259</td>
<td>0.352</td>
<td>Madagascar, 2022-04</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.307*</td>
<td>0.167</td>
<td>Madagascar, 2022-04</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>0.225</td>
<td>0.567</td>
<td>Madagascar, 2021-05</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.249</td>
<td>0.169</td>
<td>Madagascar, 2021-05</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.1</td>
<td>0.266</td>
<td>Madagascar, 2021-05</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>0.13</td>
<td>0.162</td>
<td>Madagascar, 2021-05</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>-0.545</td>
<td>0.522</td>
<td>Madagascar, 2021-05</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>0.516****</td>
<td>0.102</td>
<td>Madagascar, 2022-11</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>-0.023</td>
<td>0.091</td>
<td>Madagascar, 2022-11</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.076</td>
<td>0.14</td>
<td>Madagascar, 2022-11</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>0.451**</td>
<td>0.162</td>
<td>Madagascar, 2022-11</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.074</td>
<td>0.106</td>
<td>Madagascar, 2022-11</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>0.258****</td>
<td>0.038</td>
<td>Haiti, 2021-09</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.239****</td>
<td>0.066</td>
<td>Haiti, 2021-09</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.068</td>
<td>0.042</td>
<td>Haiti, 2021-09</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>0.180***</td>
<td>0.037</td>
<td>Haiti, 2021-09</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>-0.04</td>
<td>0.044</td>
<td>Haiti, 2021-09</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>0.03</td>
<td>0.043</td>
<td>Yemen, 2020-10</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.063</td>
<td>0.043</td>
<td>Yemen, 2020-10</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.086**</td>
<td>0.041</td>
<td>Yemen, 2020-10</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>0.216***</td>
<td>0.032</td>
<td>Yemen, 2020-10</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.209****</td>
<td>0.049</td>
<td>Yemen, 2020-10</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>0.125*</td>
<td>0.069</td>
<td>Congo, 2021-09</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.227****</td>
<td>0.078</td>
<td>Congo, 2021-09</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.092</td>
<td>0.058</td>
<td>Congo, 2021-09</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>-0.103</td>
<td>0.07</td>
<td>Congo, 2021-09</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.061</td>
<td>0.042</td>
<td>Congo, 2021-09</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>-0.084</td>
<td>0.098</td>
<td>Afghanistan, 2022-03</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.145</td>
<td>0.106</td>
<td>Afghanistan, 2022-03</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.109</td>
<td>0.069</td>
<td>Afghanistan, 2022-03</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>-0.052</td>
<td>0.056</td>
<td>Afghanistan, 2022-03</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.120**</td>
<td>0.052</td>
<td>Afghanistan, 2022-03</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>0.215**</td>
<td>0.083</td>
<td>Afghanistan, 2022-09</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------</td>
<td>-------</td>
<td>----------------------</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.097</td>
<td>0.09</td>
<td>Afghanistan, 2022-09</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>-0.102</td>
<td>0.089</td>
<td>Afghanistan, 2022-09</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>0.143***</td>
<td>0.065</td>
<td>Afghanistan, 2022-09</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.074*</td>
<td>0.042</td>
<td>Afghanistan, 2022-09</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>-0.037</td>
<td>0.078</td>
<td>Afghanistan, 2021-03</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>-0.7</td>
<td>0.086</td>
<td>Afghanistan, 2021-03</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.097</td>
<td>0.091</td>
<td>Afghanistan, 2021-03</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>0.130**</td>
<td>0.055</td>
<td>Afghanistan, 2021-03</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.288***</td>
<td>0.075</td>
<td>Afghanistan, 2021-03</td>
</tr>
<tr>
<td>FCS 3+ population (%)</td>
<td>-0.133</td>
<td>0.255</td>
<td>Afghanistan, 2021-09</td>
</tr>
<tr>
<td>rCSI 3+ population (%)</td>
<td>0.099</td>
<td>0.168</td>
<td>Afghanistan, 2021-09</td>
</tr>
<tr>
<td>HHS 3+ population (%)</td>
<td>0.004</td>
<td>0.15</td>
<td>Afghanistan, 2021-09</td>
</tr>
<tr>
<td>LCS 3+ population (%)</td>
<td>0.017</td>
<td>0.09</td>
<td>Afghanistan, 2021-09</td>
</tr>
<tr>
<td>HDDS 3+ population (%)</td>
<td>0.479*</td>
<td>0.28</td>
<td>Afghanistan, 2021-09</td>
</tr>
</tbody>
</table>

Coefficients and standard errors for each FSI are calculated using TWG analysis level OLS regressions to predict the population (%) in IPC Phase 3+. We exclude results for constants. Asterisks ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively (n=599).

Note: ‘Congo’ denotes Democratic Republic of Congo.
ANNEX 4: Bibliography


