Received 9 October 2024; accepted 27 October 2024. Date of publication 29 October 2024; date of current version 11 November 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3487912

Subgroup-Centric Multicast Cell-Free Massive MIMO

ALEJANDRO DE LA FUENTE [1], GUILLEM FEMENIAS [2] (Senior Member, IEEE), FELIP RIERA-PALOU [2] (Senior Member, IEEE), AND GIOVANNI INTERDONATO [3] (Member, IEEE)

[1] Department of Signal Theory and Communications, Universidad Rey Juan Carlos, 28942 Fuenlabrada, Spain
[2] Mobile Communications Group, University of the Balearic Islands, 07122 Palma, Spain
[3] Department of Electrical and Information Engineering, University of Cassino and Southern Lazio, 03043 Cassino, Italy

CORRESPONDING AUTHOR: A. DE LA FUENTE (e-mail: alejandro.fuente@urjc.es)

This work was supported in part by IRENE-STARMAN funded by MCIN/AEI/10.13039/501100011033, Spain, under Grant PID2020-115323RB-C32; in part by GERMINAL funded by MCIN/AEI/10.13039/501100011033, Spain, and by European Union “NextGenerationEU”/PRTR under Grant TED2021-131624B-I00; in part by ANTHEM5G funded by MCIN/AEI/10.13039/501100011033, Spain, and by European Union “NextGenerationEU”/PRTR under Grant TED2021-131975A-I00; in part by POLIGRAPH funded by MCIN/AEI/10.13039/501100011033, Spain, and by European Union “NextGenerationEU”/PRTR under Grant PID2022-136887NB-I00; in part by SOFIA-WIND funded by MICIU/AEI/10.13039/501100011033 and FEDER, EU under Grant PID2023-147305OB-C33; and in part by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (Program “RESTART,” Structural Project 6GWINET, Cascade Call SPARKS, CUP D43C22003080001) under Grant PE00000001.

ABSTRACT Cell-free massive multiple-input multiple-output (CF-mMIMO) is an emerging technology for beyond fifth-generation (5G) systems aimed at enhancing the energy and spectral efficiencies of future mobile networks while providing nearly uniform quality of service to all users. Moreover, multicasting has garnered increasing attention in recent years, as physical-layer multicasting proves to be an efficient approach for serving multiple users simultaneously, all with identical service demands while sharing radio resources. A multicast service is typically delivered using either unicast or a single multicast transmission. In contrast, this work introduces a subgroup-centric multicast CF-mMIMO framework that splits the users into several multicast subgroups. The subgroup creation is based on the similarities in the spatial channel characteristics of the multicast users. This framework benefits from efficiently sharing the pilot sequence used for channel estimation and the precoding filters used for data transmission. The proposed framework relies on two scalable precoding strategies, namely, the centralized improved partial MMSE (IP-MMSE) and the distributed conjugate beamforming (CB). Numerical results demonstrate that the centralized IP-MMSE precoding strategy outperforms the CB precoding scheme in terms of sum SE when multicast users are uniformly distributed across the service area. In contrast, in cases where users are spatially clustered, multicast subgrouping significantly enhances the sum spectral efficiency (SE) of the multicast service compared to both unicast and single multicast transmission. Interestingly, in the latter scenario, distributed CB precoding outperforms IP-MMSE, particularly in terms of per-user SE, making it the best solution for delivering multicast content. Heterogeneous scenarios that combine uniform and clustered distributions of users validate multicast subgrouping as the most effective solution for improving both the sum and per-user SE of a multicast CF-mMIMO service.

INDEX TERMS Cell-free massive MIMO, multicasting, user subgrouping, scalability.

I. INTRODUCTION

GLOBAL mobile traffic has grown unprecedentedly over the last decade, reaching 130 EB/month by 2023, with no signs of slowing down. On the contrary, fueled by the demands of emerging applications, current predictions envisage that mobile data traffic will triple, reaching an estimated 403 EB/month by 2029 [1]. This sheer volume of traffic poses enormous challenges to mobile operators and manufacturers in terms of spectral and energy efficiencies, which the current fifth-generation (5G) wireless communications standard is unlikely to meet. Consequently, academia, industry, and standardization bodies are pushing forward the sixth-generation (6G) wireless communications standard likely to debut in the early 2030s [2]. Although discussions are ongoing regarding the innovation pillars supporting 6G, some technologies are expected to play a chief role. Notable among these are new forms of massive multiple-input multiple-output (mMIMO), the integration of terrestrial and non-terrestrial segments, and the use of new frequency bands (e.g., centimeter-wave (cmWave) commonly referred to as FR3, THz), which are often heralded as the enablers of the 5G-to-6G transition [2], [3].

Cell-free mMIMO (CF-mMIMO), first reported in [4], is the evolution of mMIMO that has recently gathered the most interest from the research community. In a CF-mMIMO system, many access points (APs) are spread out over a coverage area and linked to one or multiple central processing units (CPUs) via fronthaul links to potentially share both control and data planes. The APs synchronously and coherently serve all active users on the same time-frequency resources and the use of advanced joint processing schemes results in unprecedented degrees of macro-diversity and spatial multiplexing gain [4], [5], [6], [7]. CF-mMIMO synergistically combines the main attributes of co-located mMIMO and network MIMO, providing users with nearly uniform high-quality service across the coverage area [4]. The outstanding enhancements that this architecture brings in terms of spectral and energy efficiency, service quality, and reliability have garnered tremendous interest from academia and industry in recent years [3], [8], [9]. Driven by the promising results of the seminal work [4], where a tenfold improvement in the 95%-likely spectral efficiency (SE) was demonstrated when compared to a small-cell system, intensive research efforts have been carried out over the last few years to fully characterize its performance [4], [10], [11], [12], [13]. Specifically, Björnson and Sanguinetti in [13] extended the framework to spatially correlated Rayleigh fading channels and different levels of cooperation in the uplink (UL). The authors in [6] investigated low-complexity hybrid precoders/decoders for CF-mMIMO systems operating at millimeter-waves (mmWaves) under the assumption of capacity-limited fronthaul links. The downlink (DL) energy efficiency (EE) of a CF-mMIMO system was maximized in [12] by relying on an accurate power consumption model that accounted for channel estimation errors, AP selection methods, hardware impairments, and fronthaul power consumption.

The vast majority of cell-free literature, like the aforementioned contributions, explore the improvements attained in the SE, EE, and coverage performance in CF-mMIMO unicast transmissions through the utilization of linear signal processing and local channel state information (CSI). Critically, within the vast volume of data traffic that has been predicted, a significant portion will comprise content that can potentially be shared among groups of users in the network, and therefore, can be leveraged through broadcast/multicast techniques [14], [15]. Multicasting, that is, the simultaneous delivery of common data to multiple users, can play a central role in group-oriented services such as live video streaming (e.g., Twitch), virtual reality, software updates, and Internet of things (IoT) applications, thus motivating the search for efficient and scalable multicast solutions in a 6G context. In particular, CF-mMIMO constitutes a promising backbone technology to address the challenges envisaged in next-generation multicast technology. Moreover, the superior energy efficiency of CF-mMIMO [16] complements the goal of minimizing unnecessary transmissions as advocated by the multicast paradigm, thus jointly contributing to a greener and more sustainable communication ecosystem. Multicasting can also improve the reliability of these services by reducing the impact of network congestion and failures. The *experience sharing* (i.e., 4K/8K high-definition (HD) video streaming, extended reality (XR), and holographic communications) includes applications demanding very high reliability and data rate as well as extremely low latency. Multicast traffic delivery can bring significant benefits to both communication and computing components in such a bandwidth-demanding scenario, especially in dense user areas. The *remote control and robotic technology* can also benefit from multicasting, which allows for software upgrades to a group of IoT devices or for switching on a set of street lights simultaneously. Furthermore, recent advancements in pervasive connection and ubiquitous computing have generated many applications demanding real-time information updates in the networks of *environmental sensors and self-driving cars*. Group-based interactions are distinguished in many pervasive computing scenarios due to their user-centric characteristics and group interaction capabilities. For example, time-sensitive data gathering and monitoring systems can minimize the age of information and save bandwidth by sending the update messages through multicast channels [17]. A clear indication of the growing importance of multicast can be found in recent standardization activities. In particular, techniques related to multimedia multicast transmission have been standardized in the 3rd Generation Partnership Project (3GPP) Release-17 of the *New Radio* specifications for 5G and beyond under the name of *evolved Multimedia Broadcast and Multicast Service (eMBMS)* [18]. Moreover, for the next step in the evolution of *New Radio eMBMS*, one of the most important enhancements, targeted for Release-18, is to extend the support of multicast to users in *inactive* state. This will be an important enhancement to better support extreme congestion, where the number of users is too large to keep all of these in *connected* state [19].

Despite the evident benefits of combining CF-mMIMO and multicast, critical issues remain. While precoding and power allocation strategies have been extensively studied in unicast CF-mMIMO scenarios, there is no direct and clear-cut translation between the most efficient configurations for unicast transmissions and those for multicast setups. In particular, unknowns persist regarding the design of precoders and the allocation of power when assuming that transmissions are shared among the service-specific group of multicast users, potentially organized into different multicast subgroups due to the widely different channel conditions each subgroup is experiencing.

Subgroup-centric multicast, as proposed in this research work, can be seen as a mechanism to tailor the transmission of shared content to subgroups of users experiencing similar intra-subgroup and widely different inter-subgroup propagation conditions. Critical issues such as the proper management of intra-subgroup and inter-subgroup pilot contamination, the design of common channel estimation processes for all users within the same multicast subgroup, and the impact these common channel estimates might have on the design of the common precoder used to convey the DL multicast payload data to users remain largely unsolved.

A. RELATED WORKS

Delivering a common data service to a set of users has traditionally been accomplished by two opposite strategies: either through a single multicast transmission to the whole set of users or through multiple unicast transmissions, each intended for a single user of the multicast group. In the first case, as there is a single transmission, there is no inter-user interference. However, to ensure that all users can properly decode the received signal, the SE of the multicast transmission is limited by the channel characteristics of the user experiencing the worst propagation conditions. In contrast, with the second strategy, the SE of each unicast transmission can be adapted to the specific propagation conditions of each user. Nevertheless, as the number of users can be very large, all received signals are typically affected by high levels of inter-user interference. To strike a proper trade-off between the advantages and disadvantages of these strategies, a multicast subgrouping approach has been already proposed and evaluated. This approach has shown potential in traditional single-input single-output (SISO), MIMO, and massive MIMO systems, using both wideband and subband CSI [20], [21], [22].

In the last few years, the proliferation of mMIMO research has led to the development of novel multi-group multicast transmission strategies for co-located mMIMO systems [22], [23], [24], [25]. These strategies are based on using a common pilot sequence for all the multicast users within the same group, allowing them to share a single channel estimate and precoder vector per multicast group. A similar multi-group multicast framework has also been applied to CF-mMIMO architectures in [26], where the authors propose a novel DL pilot training scheme and present a detailed analysis of the DL performance. In [27], the authors proposed a weighted *max-min* power optimization algorithm for multi-group multicast CF-mMIMO. Moreover, recent studies have investigated non-orthogonal multiple access (NOMA) with unicast and multicast transmissions [28], multi-antenna multicast users with low-resolution analog-digital converters (ADCs)/digital-analog converters (DACs) [29], and instantaneous power control policies in multi-group multicast CF-mMIMO [30]. The work [31] introduces efficient RIS-based methods for multi-group multicasting, highlighting a multicasting tailored zero-forcing (MTZF) beamforming technique that efficiently suppresses the inter-group interference, thereby able to provide high levels of sum rate with fewer antennas and low computational complexity.

B. MAIN CONTRIBUTIONS

Many of the aforementioned research works addressing the integration of CF-mMIMO and multicast suffer from several shortcomings, notably the utilization of simplified channel models and precoding techniques. Uncorrelated Rayleigh fading and conjugate beamforming (CB) are commonly employed in the literature, thereby prompting the approach proposed in this paper, which incorporates spatial channel correlation and explores the advantages and disadvantages of more sophisticated precoding techniques depending on the users' distribution within the CF-mMIMO-multicast framework. Moreover, past research has largely overlooked the impact of pilot contamination, and our study demonstrates its significance, particularly in scenarios where a massive number of users demand the same multicast service. Additionally, in contrast to prior studies that assume the delivery of each multicast service through a single transmission to all users, regardless of their locations and corresponding large-scale channel similarities, our work introduces a novel *subgroup-centric* framework. In this framework, a single multicast service is delivered to disjoint subgroups of users via multicast transmissions. Each multicast transmission entails that the users to which it is concerned must share the same UL pilot and DL precoding filter. Therefore, we expect that users experiencing similar large-scale channel characteristics may share these resources efficiently. Thus, our claim is that subgrouping multicast users based on large-scale propagation similarities may provide significant enhancements in spectral efficiency, resulting from more efficient use of both UL pilots and DL data resources. In summary, the main contributions of this research work are:

- The design of a novel users' subgroup-centric framework for the multicast CF-mMIMO DL that accounts for spatially correlated fading channels and pilot contamination.
- The proposal of a multicast user subgrouping method that is grounded in large-scale similarity metrics that characterize propagation channels between users and their corresponding sets of serving APs. This approach enables the design of a pilot allocation strategy aimed at reducing inter-subgroup pilot contamination, thereby enhancing the effectiveness of the subgroup transmit precoding technique.
- A comprehensive performance evaluation of two precoding schemes, along with their corresponding power control strategies, appropriately tailored to the proposed subgroup-centric multicast framework, namely, the centralized improved partial MMSE (IP-MMSE) [32] and the distributed CB, both incorporating specifically designed fractional power control strategies inspired by the proposals of Demir et al. in [33]. Notably, a closed-form expression is derived for the achievable SE in the multicast distributed CB scenario.
- Finally, comprehensive simulation results are presented to quantify the benefits of the proposed user subgrouping, precoding techniques, and power allocation

strategies across various system setups. This is achieved by benchmarking the subgroup-centric multicasting approach introduced in this research against conventional multicasting and unicast transmission strategies. This comparison takes into account both uniform and clustered spatial distributions of users.

C. PAPER OUTLINE AND NOTATIONS

The remainder of this paper is organized as follows. In Section II, the system model is described, including the proposed subgroup-centric CF-mMIMO multicasting approach over spatially correlated Rayleigh fading channels. In Section III, the multicast subgrouping framework is further explained by including a detailed description of the multicast AP cooperation clustering and pilot allocation strategies, the UL training phase, the DL payload data transmission phase, the precoding techniques, and the fractional power allocation schemes. Numerical results are presented in Section IV to assess the technical soundness of the proposed strategies. Finally, Section V concludes the paper by discussing the significance of subgrouping multicast users according to their large-scale channel similarities in CF-mMIMO scenarios, while also providing insights into potential future research directions.

Notational Remark: Lowercase and uppercase boldface letters are used to denote vectors and matrices, respectively. Calligraphic uppercase letters are used to denote sets, with $|\mathcal{A}|$ denoting the cardinality of set $\mathcal{A}$. The superscripts $(\cdot)^{\mathsf{T}}$, $(\cdot)^{*}$ and $(\cdot)^{\mathsf{H}}$ denote the transpose, conjugate and conjugate transpose (Hermitian) operators, respectively. The set of complex numbers is represented by $\mathbb{C}$. $\mathbb{E}\{\cdot\}$ denotes the expectation operator. $\mathrm{tr}(\boldsymbol{A})$ denotes the trace of matrix $\boldsymbol{A}$. A circularly symmetric complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\mathbf{I}_N$ represents the $N \times N$ identity matrix. $\|\boldsymbol{a}\|$ denotes the ℓ_2-norm of the vector $\boldsymbol{a}$.

II. SYSTEM MODEL

A CF-mMIMO system operating in time division duplexing (TDD) [4] is considered where L APs, each equipped with N antennas, are connected via ideal fronthaul links to a CPU. The APs are uniformly distributed over the coverage area and simultaneously provide a shared data service, either through multicast or unicast, to K single-antenna users on the same time-frequency resources. The set of users is denoted by $\mathcal{K}$ and indexed by $k \in \mathcal{K} = \{1, \ldots, K\}$. The set of APs is denoted by $\mathcal{L}$ and indexed by $l \in \mathcal{L} = \{1, \ldots, L\}$. Since the focus in this paper is on multicasting techniques, which always take place in the DL, we omit the study of the UL data transmission phase. Thus, we assume that each TDD frame is divided into UL training phase and DL payload data transmission phase, whose sizes, measured in samples (or channel uses), are denoted as τ_{p} and τ_{d}, respectively. The TDD frame, with size $\tau_{\mathrm{c}} = \tau_{\mathrm{p}} + \tau_{\mathrm{d}}$ samples, is assumed to fit the channel coherence block, consisting of a number of subcarriers and time samples over which the channel response is approximately frequency-flat and time invariant [34]. In practice, this multicast service can coexist with UL and DL unicast services, however, based on the available literature approaches [33], [35], we consider that omitting the UL payload phase is an adequate model to assess the subgroup-centric CF-mMIMO multicasting framework since the SE scales linearly with τ_{d} and the conclusions will be the same including a UL payload phase in our framework.

A. SUBGROUP-CENTRIC CF-MMIMO MULTICASTING

Our main aim in this work is to determine a proper trade-off between delivering a multicast service using multiple unicast transmissions and using a single multicast transmission. In this regard, a users' subgroup-centric framework is proposed where a single multicast service is delivered to $G \leq K$ disjoint subgroups of users via G multicast transmissions, assuming that there exists a one-to-one mapping between a multicast transmission and a subgroup. Each multicast transmission is characterized by a unique modulation and coding scheme (MCS) and a precoding vector, designed to match the SE requirements of the user experiencing the worst propagation conditions in the corresponding subgroup (i.e., the SE achieved by each multicast transmission is determined by the lowest signal-to-interference-plus-noise ratio (SINR) experienced by the users in the multicast subgroup). Note that carrying out CF-mMIMO multicasting, without a subgroup-centric approach leveraging the spatial distribution characteristics of the users, may possibly fade away the intrinsic feature of CF-mMIMO systems in ensuring uniformly great QoS to every user, providing, in turn, a uniformly poor service. Interestingly, in scenarios where the users are distributed in spatial clusters, subgrouping the users that experience similar large-scale channel characteristics may be beneficial. In this case, users within the same multicast subgroup share the same pilot sequence, thus allowing the estimation of a common average channel for the entire subgroup. Due to the similarity between their channel statistics, this should also provide a quite good estimation of the individual channels of each user in the subgroup. Furthermore, as they can also utilize a common DL precoder properly adapted to the shared subgroup channel characteristics of all users in the group, this can result in improved SEs.

The set of multicast subgroups is denoted by $\mathcal{G}$ and the subgroups are indexed by $g \in \mathcal{G} = \{1, \ldots, G\}$. The set of users in subgroup g is denoted by $\mathcal{K}_g$. Letting $K_g = |\mathcal{K}_g|$ be the number of users in subgroup g, it holds that $K = \sum_{g=1}^{G} K_g$. Somehow re-adapting the concept of user-centric transmission [36] and dynamic cooperation clustering (DCC) [33], a multicast subgroup-centric transmission is implemented wherein the users in multicast subgroup g are served by a subset of APs. We denote the subset of APs serving subgroup g by $\mathcal{L}_g \subseteq \{1, \ldots, L\}$, where $|\mathcal{L}_g| = L_g \leq L$. For later convenience, given a subgroup g, we define the set $\mathcal{S}_g$ as the collection of multicast subgroups served by some (or all) of the APs serving subgroup g,

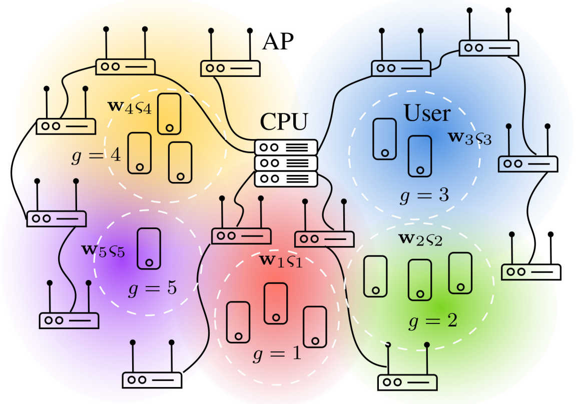

AP
CPU
User
$\mathbf{w}_4\varsigma_4$
$g=4$
$\mathbf{w}_3\varsigma_3$
$g=3$
$\mathbf{w}_5\varsigma_5$
$g=5$
$\mathbf{w}_1\varsigma_1$
$g=1$
$\mathbf{w}_2\varsigma_2$
$g=2$

FIGURE 1. A CF-mMIMO network with multicast user subgrouping (dashed circles) and AP cooperation clustering (blurred colored shapes).

that is, $\mathcal{S}_g = \{c : \mathcal{L}_g \cap \mathcal{L}_c \neq \emptyset\}$, note that $c, g \in \mathcal{G}$. The set of multicast subgroups served by AP l is denoted as $\mathcal{D}_l$. Figure 1 illustrates the system model of CF-mMIMO multicasting with user subgrouping and DCC.

B. CHANNEL MODEL

A conventional block-fading channel model is considered wherein the channel is time-invariant and frequency flat within a time-frequency coherence block and varies independently over different coherence blocks (block fading). The channel response vector $\boldsymbol{h}_{lk} \in \mathbb{C}^N$ between AP l and the multicast user k, in an arbitrary coherence block,[1] is distributed as $\boldsymbol{h}_{lk} \sim \mathcal{CN}(\mathbf{0}_N, \boldsymbol{R}_{lk})$, where $\boldsymbol{R}_{lk} \in \mathbb{C}^{N\times N}$ is the corresponding positive semi-definite spatial covariance matrix, with average channel gain given by $\beta_{lk} = \mathrm{tr}(\boldsymbol{R}_{lk})/N$. Given the reasonable physical separation between APs and users, it is justifiable to assume that the channel vectors of distinct AP-user pairs are independently distributed and experiencing uncorrelated small-scale fading, that is, $\mathbb{E}\{\boldsymbol{h}_{l'k'}\boldsymbol{h}_{lk}^{\mathsf{H}}\} = \mathbf{0}_{N\times N}, \ \forall l'k' \neq lk$. Thus, the channel from user k to the complete set of APs $l \in \mathcal{L}$, $\boldsymbol{h}_k = [\boldsymbol{h}_{1k}^T \ldots \boldsymbol{h}_{Lk}^T]^T$, is distributed as $\boldsymbol{h}_k \sim \mathcal{CN}(\mathbf{0}_{LN}, \boldsymbol{R}_k)$, where $\boldsymbol{R}_k = \mathrm{blkdiag}(\boldsymbol{R}_{1k}, \ldots, \boldsymbol{R}_{Lk}) \in \mathbb{C}^{LN\times LN}$ is the block-diagonal spatial covariance matrix related to user k, which follows a local scattering spatial correlation model for a non-line of sight (NLOS) channel between the user k and the APs $l \in \mathcal{L} = \{1, \ldots, L\}$, equipped each one with a uniform linear array (ULA) (see [34, Sec. II-F, eq. (2.23)]. The channel covariance matrices $\boldsymbol{R}_{lk}, \forall\, k \in \mathcal{K}, \ \forall\, l \in \mathcal{L}$, can be estimated at each AP over a large-scale fading time scale (i.e., over multiple coherence blocks) and thus can be safely assumed to be perfectly known at both the APs and the CPU [37]; refer to sources [37], [38], [39], [40], [41] for practical methods to estimate spatial correlation matrices.

III. MULTICAST SUBGROUPING FRAMEWORK

A. MULTICAST USERS' SUBGROUPING

The design of a single multicast transmission entails that the users to which it concerns must share the same UL pilot and DL precoding filter. Therefore, we expect that users experiencing similar large-scale channel characteristics may share these resources efficiently. Thus, subgrouping multicast users, especially in ultra-dense scenarios with users distributed in clusters, may provide enhancements in the sum SE, as a result of a more efficient use of both UL pilots and DL data resources. In [22], the authors proposed a subgrouping strategy that suits the propagation characteristics of the users in co-located mMIMO scenarios. Specifically, they considered a metric to create subgroups of users based on the degree of similarity between their channel covariance matrices. However, this strategy does not appear to be easily applicable to CF-mMIMO scenarios. In [42], the authors utilized the unique characteristics of the CF-mMIMO propagation channels to introduce a K-means-based algorithm that utilizes a simplified metric relying on the average channel gain vectors $\boldsymbol{\beta}_k = [\beta_{1k} \ldots \beta_{Lk}]^{\mathsf{T}}$, treated as an effective feature vector (fingerprint) characterizing user k, for partitioning the users into groups aimed at minimizing the effects of pilot contamination. Inspired by the aforementioned partitioning strategy, the K-means-based algorithm proposed in [42] is suitably adapted in this work to group those users experiencing similar large-scale channel characteristics into multicast subgroups that will benefit from sharing the same UL pilot sequence to improve, first, the quality of the channel estimates and, as a result, the effectiveness of the precoding technique used for the DL payload data transmission. To accomplish this, the K-means clustering algorithm is provided with the number of multicast subgroups G to be generated, along with the fingerprint vectors $\boldsymbol{\beta}_k$ of the K users in the network. The metric used to measure the distance between two fingerprint vectors is the *cosine similarity*, defined as

$$f_{\mathrm{d}}(\boldsymbol{\beta}_k, \boldsymbol{\beta}_{k'}) = \frac{\boldsymbol{\beta}_k^{\mathsf{T}}\boldsymbol{\beta}_{k'}}{\|\boldsymbol{\beta}_k\|\|\boldsymbol{\beta}_{k'}\|}. \tag{1}$$

The value of $f_{\mathrm{d}}(\boldsymbol{\beta}_k, \boldsymbol{\beta}_{k'})$ in this context varies from 0, indicating orthogonality or decorrelation, to 1, signifying identical vectors. Intermediate values denote varying degrees of similarity or dissimilarity. The output of the subgrouping algorithm is the set of clusters $\mathcal{K}_g$ for all $g \in \{1, \ldots, G\}$, where each user belongs to the subgroup with the nearest *collective* fingerprint (which represents the subgroup centroid). The rationale behind this subgroup formation strategy is that users with values of the cosine similarity between their fingerprints close to 1 are bound to be geographically close to each other, and therefore, experiencing similar statistical channel characteristics, should belong to the same multicast subgroup. By doing so, inter-subgroup pilot contamination can be mitigated, leading to improvements in channel estimation quality and precoding effectiveness [42].

[1] For the sake of clarity, we omit the index identifying the coherence block.

B. MULTICAST AP COOPERATION CLUSTERING AND PILOT ALLOCATION

The pilot sequence length is equal to τ_p samples implying that, at most, τ_p mutually orthogonal pilot sequences can be generated. Each user within a given subgroup is allocated the same pilot sequence selected from the pool of τ_p available mutually orthogonal pilots. Note that, as co-pilot users have linearly dependent channel estimates [34], [43], the APs cannot separate the users of the same subgroup in the spatial domain. Finding the optimal pilot allocation is a combinatorial optimization problem. There are τ_p^G possible assignments in a setup with G subgroups and τ_p pilots; thus, the complexity of evaluating all of them grows exponentially with the number of subgroups. Besides, the AP-to-subgroup association has to be established. To jointly address these tasks, we design a slight variant of the joint DCC and pilot allocation algorithm proposed in [33], which is applied to multicast subgroups rather than single users. The proposed joint DCC and multicast subgroup pilot allocation scheme, the pseudocode of which is presented in Alg. 1, iteratively assigns pilots to subgroups by selecting in each iteration the one leading to the minimum pilot contamination. To that end, pilots are first assigned to subgroups, and then each AP is allowed to serve exactly τ_p subgroups. For every pilot, the AP serves the subgroup with the strongest common average channel gain, i.e., $\frac{1}{K_g}\sum_{k\in\mathcal{K}_g}\beta_{lk}$, among the set of subgroups that have been assigned that pilot. As shown in Algorithm 1, the multicast subgroup pilot assignment and cooperation clustering creation process consists of two steps. In the first step, a maximum of τ_p subgroups arbitrarily indexed from 1 to $\min(\tau_\mathrm{p}, G)$ are assigned mutually orthogonal pilots, that is, every user $k \in \mathcal{K}_g$ uses pilot g for $g \in \{1, \ldots, \min(\tau_\mathrm{p}, G)\}$. If $G > \tau_\mathrm{p}$, the remaining subgroups, with indices ranging from $\tau_\mathrm{p} + 1$ to G, are then assigned pilots one after the other as follows. Following a similar rationale as in [33], assuming that AP l presents a good average large-scale propagation gain for users in subgroup g, it is expected to contribute significantly to the service quality of this particular subgroup. Consequently, it is preferable to assign subgroup g to the pilot for which AP l experiences the least pilot contamination. Hence, for each pilot $\psi \in \{1, \ldots, \tau_\mathrm{p}\}$, AP l computes the sum of the channel gains β_{lk} of the users $k \in \mathcal{K}_c,\ c \neq g,\ \boldsymbol{\psi}_c = \boldsymbol{\psi}$, that is the users that have already been allocated this pilot, and then identifies the index of the pilot minimizing the expected pilot interference as

$$\tau \leftarrow \arg\min_{\boldsymbol{\psi}\in\{1,\ldots,\tau_\mathrm{p}\}} \sum_{\substack{c\in\mathcal{G}\backslash\{g\}\\ \boldsymbol{\psi}_c=\boldsymbol{\psi}}} \sum_{k\in\mathcal{K}_c} \beta_{lk}. \qquad (2)$$

Pilot τ is then assigned to subgroup g and the algorithm continues with the next subgroup.

In the second step of the algorithm, the clusters of APs are created after all the subgroups have been assigned to pilots. Each AP evaluates, for each pilot sequence, which subgroup experiences the largest common average channel gain among those using a specific pilot sequence. The resulting subgroup is then picked to be served by this specific AP. To guarantee the service to every subgroup, a multicast subgroup g is always served by at least its own *master* AP l, which is selected by the CPU as $l = \arg\max_{l'\in\mathcal{L}}\frac{1}{K_g}\sum_{k\in\mathcal{K}_g}\beta_{l'k}$.

Algorithm 1 Multicast Subgroup Pilot Assignment and Cooperation Clustering

```
Initialization:
$\mathcal{L}_g = \emptyset \quad \forall g \in \{1,\ldots,G\}$
Input: $\tau_\mathrm{p}, \beta_{lk}, G, \mathcal{K}_g$
for $g = 1,\ldots,\min(\tau_\mathrm{p}, G)$ do
    $\boldsymbol{\psi}_g \leftarrow g$
end for
if $G > \tau_\mathrm{p}$ then
    for $g = \tau_\mathrm{p}+1,\ldots, G$ do
        $l \leftarrow \arg\max_{l\in\mathcal{L}}\frac{1}{K_g}\sum_{k\in\mathcal{K}_g}\beta_{lk}$
        $\tau \leftarrow \arg\min_{\boldsymbol{\psi}\in\{1,\ldots,\tau_\mathrm{p}\}} \sum_{c\in\mathcal{G}\backslash\{g\},\ \boldsymbol{\psi}_c=\boldsymbol{\psi}} \sum_{k\in\mathcal{K}_c}\beta_{lk}$
        $\boldsymbol{\psi}_g \leftarrow \tau$
    end for
end if
for $l = 1,\ldots,L$ do
    for $\boldsymbol{\psi} = 1,\ldots,\tau_\mathrm{p}$ do
        $c \leftarrow \arg\max_{g\in\{1,\ldots,G\}:\boldsymbol{\psi}_g=\boldsymbol{\psi}}\frac{1}{K_g}\sum_{k\in\mathcal{K}_g}\beta_{lk}$
        $\mathcal{L}_c \leftarrow \mathcal{L}_c \cup \{l\}$
    end for
end for
for $g = 1,\ldots,G$ do
    if $\mathcal{L}_g = \{\emptyset\}$ then
        $l \leftarrow \arg\max_{l\in\mathcal{L}}\frac{1}{K_g}\sum_{k\in\mathcal{K}_g}\beta_{lk}$
        $\mathcal{L}_g \leftarrow \mathcal{L}_g \cup \{l\}$
    end if
end for
Output: Pilot assignment $\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_G$ and DCCs $\mathcal{L}_1,\ldots,\mathcal{L}_G$
```

C. UPLINK SUBGROUP CHANNEL ESTIMATION

Let $\boldsymbol{\psi}_g \in \mathbb{C}^{\tau_\mathrm{p}}$ be the pilot sequence assigned to subgroup g, with $\|\boldsymbol{\psi}_g\|^2 = 1$. Ideally, pilot sequences allocated to different subgroups should be mutually orthogonal. In practical scenarios, however, it often holds that $G > \tau_\mathrm{p}$, and a given pilot sequence may be assigned to more than one subgroup, thus resulting in the pilot contamination phenomenon. The $N \times \tau_\mathrm{p}$ UL received pilot signal matrix at AP l is

$$\boldsymbol{Y}_l = \sqrt{\tau_\mathrm{p} P_\mathrm{p}} \sum_{g=1}^{G} \sum_{k\in\mathcal{K}_g} \boldsymbol{h}_{lk}\boldsymbol{\psi}_g^\mathsf{T} + \boldsymbol{N}_l, \qquad (3)$$

where P_p is the transmit power per pilot-symbol, assumed to be the same for all the users, and $\boldsymbol{N}_l \in \mathbb{C}^{N\times\tau_\mathrm{p}}$ is the additive white Gaussian noise (AWGN) matrix at AP l, whose elements are independent and identically distributed as $\mathcal{CN}(0, \sigma_\mathrm{u}^2)$. To estimate the channel of users in subgroup

g, the received UL training signal is projected onto $\boldsymbol{\psi}_g^*$ to obtain

$$\boldsymbol{y}_l^g = \sqrt{\tau_\mathrm{p} P_\mathrm{p}} \sum_{k\in\mathcal{K}_g} \boldsymbol{h}_{lk} + \sqrt{\tau_\mathrm{p} P_\mathrm{p}} \sum_{\substack{c\in\mathcal{G}\setminus\{g\} \\ \boldsymbol{\psi}_c=\boldsymbol{\psi}_g}} \sum_{i\in\mathcal{K}_c} \boldsymbol{h}_{li} + \boldsymbol{n}_{lg}, \quad (4)$$

where $\boldsymbol{n}_{lg} = \boldsymbol{N}_l \boldsymbol{\psi}_g^* \sim \mathcal{CN}(0, \sigma_\mathrm{u}^2 \mathbf{I}_N)$.

Since the APs cannot separate co-pilot users in the spatial domain as their channel estimates are correlated, we define the subgroup channel of the users in $\mathcal{K}_g$ as

$$\boldsymbol{h}_l^g = \frac{\sqrt{\tau_\mathrm{p} P_\mathrm{p}}}{K_g} \sum_{k\in\mathcal{K}_g} \boldsymbol{h}_{lk}, \quad (5)$$

which is distributed as $\boldsymbol{h}_l^g \sim \mathcal{CN}(\mathbf{0}_N, \boldsymbol{R}_l^g)$, where

$$\boldsymbol{R}_l^g = \frac{\tau_\mathrm{p} P_\mathrm{p}}{K_g^2} \sum_{k\in\mathcal{K}_g} \boldsymbol{R}_{lk}. \quad (6)$$

Lemma 1: The minimum mean-squared error (MMSE) estimate of the subgroup channel $\boldsymbol{h}_l^g$ can be obtained either at the lth AP or at the CPU (in cases where the APs are not equipped with local baseband processors) as [34, Sec. 3.2]

$$\hat{\boldsymbol{h}}_l^g = K_g \, \boldsymbol{R}_l^g \, \boldsymbol{\Gamma}_{lg}^{-1} \, \boldsymbol{y}_l^g \quad (7)$$

where

$$\boldsymbol{\Gamma}_{lg} = \tau_\mathrm{p} P_\mathrm{p} \sum_{\substack{c\in\mathcal{G} \\ \boldsymbol{\psi}_c=\boldsymbol{\psi}_g}} \sum_{i\in\mathcal{K}_c} \boldsymbol{R}_{li} + \sigma_\mathrm{u}^2 \boldsymbol{I}_N. \quad (8)$$

The subgroup channel estimate is distributed as $\hat{\boldsymbol{h}}_l^g \sim \mathcal{CN}(\mathbf{0}_N, K_g^2 \boldsymbol{R}_l^g \boldsymbol{\Gamma}_{lg}^{-1} \boldsymbol{R}_l^g)$, and it is uncorrelated to the subgroup channel estimation error $\tilde{\boldsymbol{h}}_l^g \sim \mathcal{CN}(\mathbf{0}_N, \boldsymbol{R}_l^g - K_g^2 \boldsymbol{R}_l^g \boldsymbol{\Gamma}_{lg}^{-1} \boldsymbol{R}_l^g)$.

Proof: See Appendix A. ■

D. MULTICAST TRANSMISSION AND PER-SUBGROUP SE

In the proposed framework, the DL data transmission is subgroup-centric rather than user-centric (unicast) or group-centric (conventional multicast). Although all users require the same payload data content and could be easily served with a single multicast transmission, we propose instead to effectively partition them by grouping together those users experiencing similar statistical propagation conditions. Therefore, it seems logical to send a separate data stream to each of these groups, employing as many multicast transmissions as the number of subgroups, G. In this case, the DL signal received by user k of subgroup g is

$$y_k = \sum_{l=1}^{L} \boldsymbol{h}_{lk}^\mathrm{H} \boldsymbol{D}_{lg} \mathbf{w}_{lg} \varsigma_g + \sum_{l=1}^{L} \sum_{\substack{c=1 \\ c\neq g}}^{G} \boldsymbol{h}_{lk}^\mathrm{H} \boldsymbol{D}_{lc} \mathbf{w}_{lc} \varsigma_c + n_k, \quad (9)$$

where $n_k \sim \mathcal{CN}(0, \sigma_\mathrm{d}^2)$ is the AWGN at user k, $\mathbf{w}_{lg} \in \mathbb{C}^N$ represents the precoding vector used by AP l to send the multicast data to subgroup g, and ς_g denotes the data symbol intended for all users in subgroup g, with $\mathbb{E}\{|\varsigma_g|^2\} = 1$, and $\mathbb{E}\{\varsigma_g \varsigma_c^*\} = 0$, $\forall\ g \neq c$ (each multicast transmission is characterized by a unique MCS which makes the data intended to different subgroups statistically uncorrelated [22]). The set of auxiliary diagonal matrices $\boldsymbol{D}_{lg} \in \mathbb{C}^{N\times N}$ are used to represent the APs-to-subgroup association, and are given by

$$\boldsymbol{D}_{lg} = \begin{cases} \boldsymbol{I}_N, & \text{if } \ l \in \mathcal{L}_g \\ \mathbf{0}_{N\times N}, & \text{otherwise}, \end{cases} \quad (10)$$

for all $l \in \{1, \ldots, L\}$ and $g \in \{1, \ldots, G\}$. The first term in (9) denotes the desired signal, whereas the second term is the inter-subgroup interference. As conventionally assumed in mMIMO operation, the users do not acquire the DL CSI, but rely on a mean value approximation of their DL precoded channels. For the sake of brevity, let us define

$$\varrho_{lk}^{gc} = \boldsymbol{h}_{lk}^\mathrm{H} \boldsymbol{D}_{lc} \mathbf{w}_{lc} \quad k \in \mathcal{K}_g, \quad (11)$$

as the component of the DL effective channel from AP l to user k of subgroup g precoded for subgroup c. Using this definition, and leveraging channel hardening, the knowledge of statistical CSI in the form of the expected sum $\sum_{l=1}^{L} \mathbb{E}\{\varrho_{lk}^{gg}\}$ at user k in subgroup g can be considered to be a very good approximation to the instantaneous CSI $\sum_{l=1}^{L} \varrho_{lk}^{gg}$. Consequently, an accurate achievable DL SE can be obtained by applying the popular *hardening bound technique* [34], [43], [44], [45] to the signal model in (9) and treating all the interference sources as uncorrelated noise. Specifically, an achievable DL SE of an arbitrary user k in subgroup g, is given by

$$\xi_k = \left(1 - \tau_\mathrm{p}/\tau_\mathrm{c}\right) \log_2(1 + \gamma_k), \quad (12)$$

where γ_k is the effective SINR that can be expressed as

$$\gamma_k = \frac{\left|\sum_{l=1}^{L} \mathbb{E}\{\varrho_{lk}^{gg}\}\right|^2}{\sum_{c=1}^{G} \mathbb{E}\left\{\left|\sum_{l=1}^{L} \varrho_{lk}^{gc}\right|^2\right\} - \left|\sum_{l=1}^{L} \mathbb{E}\{\varrho_{lk}^{gg}\}\right|^2 + \sigma_\mathrm{d}^2}, \quad (13)$$

and the expectations are taken with respect to the random channel realizations. The expression of the achievable SE in (12) applies to any precoding scheme, multicast DCC approach, channel estimator, and channel distribution. Since each subgroup of users is served using a single multicast transmission, the achievable SE of subgroup g (and consequently the SE achievable by all users within this particular subgroup) is limited by that achievable by the worst user in the subgroup, and hence can be expressed as

$$\Xi_g = \min_{k\in\mathcal{K}_g} \xi_k. \quad (14)$$

E. SUBGROUP-CENTRIC PRECODING AND POWER ALLOCATION

In this section, two precoding schemes appropriately tailored to the proposed subgroup-centric multicast framework are considered, each accompanied by its corresponding power control strategy: the centralized IP-MMSE [32], as well as the distributed CB [4].

1) CENTRALIZED IP-MMSE PRECODING

In a centralized DL operation, all the subgroup channel estimates are available at the CPU and can be used to design the centralized precoding vectors. Let us define the collective vector of the channel estimates per subgroup as $\check{\boldsymbol{h}}^g = \boldsymbol{D}_g\hat{\boldsymbol{h}}^g$, where $\hat{\boldsymbol{h}}^g = [(\hat{\boldsymbol{h}}_1^g)^\mathsf{T}\dots(\hat{\boldsymbol{h}}_L^g)^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{LN}$, $\mathbf{w}_g = [\mathbf{w}_{1g}^\mathsf{T}\dots\mathbf{w}_{Lg}^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{LN}$ is the collective precoding vector assigned to subgroup g, and $\boldsymbol{D}_g = \text{blkdiag}(\boldsymbol{D}_{1g},\dots,\boldsymbol{D}_{Lg}) \in \mathbb{C}^{LN\times LN}$. Using this definition, the received signal in (9) can be written in compact form as

$$y_k = \boldsymbol{h}_k^\mathsf{H}\boldsymbol{D}_g\mathbf{w}_g\varsigma_g + \sum_{\substack{c=1\\c\neq g}}^{G}\boldsymbol{h}_k^\mathsf{H}\boldsymbol{D}_c\mathbf{w}_c\varsigma_c + n_k, \tag{15}$$

where $\boldsymbol{h}^g = [(\boldsymbol{h}_1^g)^\mathsf{T}\dots(\boldsymbol{h}_L^g)^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{LN}$ is the collective channel vector. This system model formulation leads to rewrite the effective SINR expression (yet equivalent to (13)) as

$$\gamma_k = \frac{\left|\mathbb{E}\{\varrho_k^{gg}\}\right|^2}{\sum_{c=1}^{G}\mathbb{E}\{\left|\varrho_k^{gc}\right|^2\} - \left|\mathbb{E}\{\varrho_k^{gg}\}\right|^2 + \sigma_\mathrm{d}^2}, \tag{16}$$

where $\varrho_k^{gc} = \boldsymbol{h}_k^\mathsf{H}\boldsymbol{D}_c\mathbf{w}_c$, for all $k \in \mathcal{K}_g$. The centralized precoding vector used to multicast data to users in subgroup g is designed, capitalizing on the UL-DL duality theorem [33], as

$$\mathbf{w}_g = \sqrt{\rho_g}\frac{\bar{\mathbf{w}}_g}{\sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_g\|^2\}}}, \tag{17}$$

where $\rho_g \geq 0$ is the total transmit power allocated to subgroup g from all the serving APs, and $\bar{\mathbf{w}}_g = [\bar{\mathbf{w}}_{1g}^\mathsf{T}\dots\bar{\mathbf{w}}_{Lg}^\mathsf{T}]^\mathsf{T}$ is the dual collective *virtual multicast*[2] combining vector. The normalization in (17) ensures that $\mathbb{E}\{\|\mathbf{w}_g\|^2\} = \rho_g$.

Since all the users belonging to subgroup g employ the same pilot, the CPU can easily exploit the knowledge of the collective subgroup channel estimates $\check{\boldsymbol{h}}^c$, for all $c \in \mathcal{S}_g$, to design the centralized precoding vector. In this regard, the IP-MMSE combining scheme [32] can be extended to the proposed multicast users' subgrouping framework by designing the collective combiner vector as

$$\bar{\mathbf{w}}_g = \sqrt{\Delta_g}\left(\sum_{c\in\mathcal{S}_g}\Delta_c\check{\boldsymbol{h}}^c\check{\boldsymbol{h}}^{c\mathsf{H}} + \boldsymbol{Z}_{\mathcal{S}_g} + \sigma_\mathrm{u}^2\boldsymbol{I}_{L_gN}\right)^{-1}\check{\boldsymbol{h}}^g, \tag{18}$$

where $\Delta_g = \frac{p_gK_g^2}{\tau_\mathrm{p}P_\mathrm{p}}$, p_g denotes the total amount of power that would be allocated to users in subgroup g in a *virtual* UL payload transmission phase, and

$$\boldsymbol{Z}_{\mathcal{S}_g} = \sum_{c\in\mathcal{S}_g}\Delta_c\boldsymbol{D}_g\tilde{\boldsymbol{R}}^c\boldsymbol{D}_g + \sum_{c\notin\mathcal{S}_g}\Delta_c\boldsymbol{D}_g\boldsymbol{R}^c\boldsymbol{D}_g, \tag{19}$$

where $\tilde{\boldsymbol{R}}^c = \text{blkdiag}(\tilde{\boldsymbol{R}}_1^c,\dots,\tilde{\boldsymbol{R}}_L^c) \in \mathbb{C}^{LN\times LN}$ denotes the error correlation matrix of the collective channel $\hat{\boldsymbol{h}}^c$ and $\boldsymbol{R}^c = \text{blkdiag}(\boldsymbol{R}_1^c,\dots,\boldsymbol{R}_L^c) \in \mathbb{C}^{LN\times LN}$ is the covariance matrix characterizing the collective channels of the interfering subgroups.

As for the power allocation strategy, an *inter-subgroup* fractional DL power control is proposed that consists in setting the power coefficient intended for subgroup g as

$$\rho_g = P_\mathrm{dl}\frac{\left[\sum_{l\in\mathcal{L}_g}\text{tr}(\boldsymbol{R}_l^g)\right]^\nu\omega_g^{-\kappa}}{\max\limits_{\ell\in\mathcal{L}_g}\sum_{c\in\mathcal{D}_\ell}\left[\sum_{l\in\mathcal{L}_c}\text{tr}(\boldsymbol{R}_l^g)\right]^\nu\omega_c^{1-\kappa}}, \tag{20}$$

where P_dl is the maximum transmit power at the APs. The parameter $\nu \in [-1, 1]$ is used to set the power allocation policy (i.e., $\nu < 0$ strives for user fairness, and $\nu > 0$ aims at sum-rate maximization). Moreover,

$$\omega_g = \max_{l\in\mathcal{L}_g}\frac{\mathbb{E}\{\|\bar{\mathbf{w}}_{lg}\|^2\}}{\mathbb{E}\{\|\bar{\mathbf{w}}_g\|^2\}}, \tag{21}$$

defined as the largest fraction of ρ_g that can be used at any of the serving APs, is used as an additional tuning parameter with an exponent $0 \leq \kappa \leq 1$ that reshapes the ratio of power allocation between different subgroups. Note that this power control strategy guarantees that [33]

$$\sum_{g\in\mathcal{D}_l}\rho_g\frac{\mathbb{E}\{\|\bar{\mathbf{w}}_{lg}\|^2\}}{\mathbb{E}\{\|\bar{\mathbf{w}}_g\|^2\}} \leq P_\mathrm{dl}. \tag{22}$$

2) DISTRIBUTED PRECODING

Distributed DL operation reduces the computational burden on the CPU by allowing most baseband processing to be conducted at local AP processors, albeit at the cost of some performance degradation. The distributed precoding vector used by AP l to multicast data to the users in subgroup g is given by

$$\mathbf{w}_{lg} = \sqrt{\rho_{lg}}\frac{\bar{\mathbf{w}}_{lg}}{\sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_{lg}\|^2\}}}, \tag{23}$$

where $\rho_{lg} \geq 0$ denotes the DL transmit power allocated to subgroup g by AP l. The CB *virtual* UL combiner avoids matrix inversions and is simply obtained by setting $\bar{\mathbf{w}}_{lg} = \boldsymbol{D}_{lg}\hat{\boldsymbol{h}}_l^g$.

A fractional DL power allocation [11], [33], [46] is considered for the distributed operation as well,

$$\rho_{lg} = \begin{cases} P_\mathrm{dl}\dfrac{\left[\text{tr}(\boldsymbol{R}_l^g)\right]^\nu}{\sum_{g\in\mathcal{D}_l}\left[\text{tr}(\boldsymbol{R}_l^g)\right]^\nu} & \text{if } g \in \mathcal{D}_l \\ 0, & \text{otherwise.}\end{cases} \tag{24}$$

Note that, in this case, the per-AP power constraint is fulfilled with equality $\forall l \in \mathcal{L}$ as $\sum_{g\in\mathcal{D}_l}\rho_{lg} = P_\mathrm{dl}$.

Lemma 2: Using the distributed CB precoder, the effective SINR in (13) can be computed in closed form[3] using

[2] The term "virtual multicast" is used as the UL traffic is of unicast type, but we apply the UL-DL duality theorem on the multicast subgroup channels defined in (5).

[3] Analytical expression results derived in Lemma 2 using 250 snapshots has been compared with the Monte Carlo simulation results employing 500 fast-fading channel realizations per snapshot, and both results match perfectly. Hence, only analytical (or Monte Carlo) results are included in the figures of this manuscript.

$$\mathbb{E}\left\{\|\boldsymbol{D}_{lg}\hat{\boldsymbol{h}}_l^g\|^2\right\} = K_g^2 \,\mathrm{tr}\left(\boldsymbol{\Lambda}_l^g \boldsymbol{R}_l^g\right), \tag{25}$$

$$\sum_{l=1}^{L} \mathbb{E}\{\varrho_{lk}^{gg}\} = \sqrt{\tau_{\mathrm{p}} P_{\mathrm{p}}} \sum_{l=1}^{L} \sqrt{\rho_{lg}} \frac{\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^g\right)}{\sqrt{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^g \boldsymbol{R}_l^g\right)}}, \tag{26}$$

$$\mathbb{E}\left\{\left|\sum_{l=1}^{L} \varrho_{lk}^{gc}\right|^2\right\} = \sum_{l=1}^{L} \rho_{lc} \frac{\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c \boldsymbol{R}_l^c \boldsymbol{D}_{lc}\right)}{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^c \boldsymbol{R}_l^c\right)} + \varsigma_{lk}^{gc}, \tag{27}$$

where $\boldsymbol{\Lambda}_l^g = \boldsymbol{D}_{lg}\boldsymbol{R}_l^g \boldsymbol{\Gamma}_{lg}^{-1}$, and

$$\varsigma_{lk}^{gc} = \begin{cases} \tau_{\mathrm{p}} P_{\mathrm{p}} \left| \sum_{l=1}^{L} \sqrt{\rho_{lc}} \frac{\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\right)}{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^c \boldsymbol{R}_l^c\right)} \right|^2, & \text{if } \ \boldsymbol{\psi}_c = \boldsymbol{\psi}_g, \\ 0, & \text{if } \boldsymbol{\psi}_c \neq \boldsymbol{\psi}_g. \end{cases} \tag{28}$$

Proof: See Appendix B. ■

F. COMPUTATIONAL COMPLEXITY ANALYSIS

The complexity of computing the precoding vectors might be an important factor to consider in selecting the adequate precoding strategy in high mobility scenarios, where the typical channel coherence interval is between 200-500 ms. In this subsection, we analyze the variables that affect the number of operations required to calculate the centralized IP-MMSE and the distributed CB precoding vectors. As we use scalable CF-mMIMO solutions, the computational complexity per subgroup g is independent of G. The simplest solution is the use of distributed CB whose corresponding precoding vector follows directly from the channel estimates. The total number of complex multiplications is $(N\tau_{\mathrm{p}} + N^2)|\mathcal{L}_g|$ per multicast subgroup (see [33, Table V.1]).

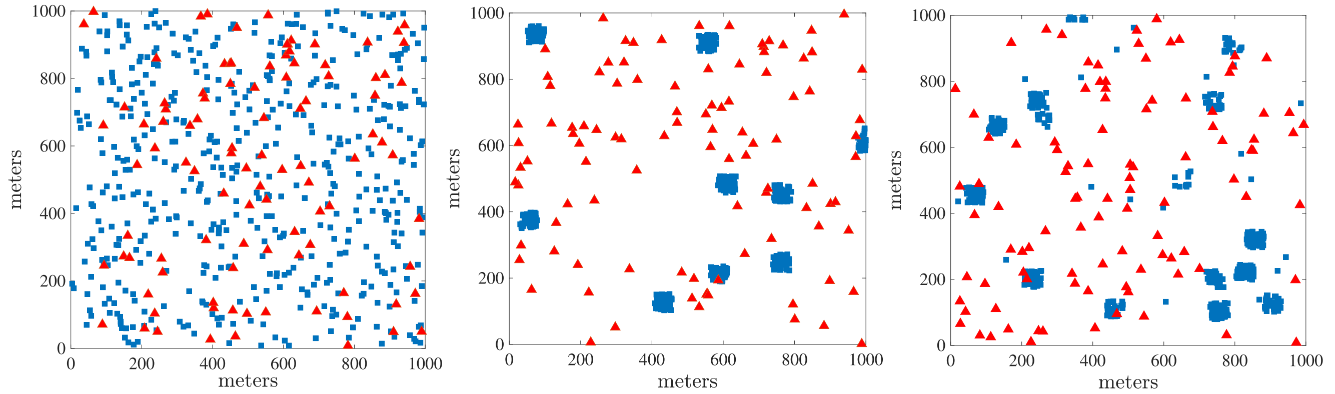
To compute the centralized IP-MMSE precoding vector per subgroup g the CPU needs to calculate the G MMSE composite channel estimates corresponding to any AP l serving subgroup g (i.e., APs with index $l \in \mathcal{L}_g$). Note that only the large-scale CSI (i.e., spatial correlation matrices) corresponding to subgroups $c \in \mathcal{S}_g$ is required at the CPU, thus, the total number of subgroups included in the inverse matrix in (18) are those served by partially the same APs as subgroup g (i.e., $|\mathcal{S}_g|$). Consequently, the total number of complex multiplications required for channel estimation is $(N\tau_{\mathrm{p}} + N^2)|\mathcal{S}_g||\mathcal{L}_g|$. Additionally, we need to account for the complexity of computing the precoding vector for each subgroup $g = 1, \ldots, G$ once per coherence block. Using the framework in [34, App. B.1.1] and considering that only a subset of the APs takes part in estimating the transmitted signal s_g, the total number of complex multiplications required for the computation of the IP-MMSE precoding vector of a subgroup g is $\frac{(N|\mathcal{L}_g|)^2 + N|\mathcal{L}_g|}{2}|\mathcal{S}_g| + (N|\mathcal{L}_g|)^2 + \frac{(N|\mathcal{L}_g|)^3 - N|\mathcal{L}_g|}{3}$.

We should emphasize that the total number of operations carried out in distributed CB precoding should be made at the AP per subgroup served by such AP at each coherence interval. On the other hand, the CPU should carry out the total number of operations required in the centralized IP-MMSE precoding per subgroup served by the APs connected to the CPU at each coherence interval. As it is well known, the computational complexity of distributed CB precoding is substantially lower than the centralized IP-MMSE precoding. Furthermore, an interesting insight that we can observe is that the fewer multicast subgroups, the lower the computational complexity. Therefore, distributed CB precoding, which performs significantly better than centralized IP-MMSE precoding in scenarios with a few multicast subgroups with a large number of users per subgroup, results in an extremely lower computational complexity solution than using unicast transmissions with centralized IP-MMSE precoding.

IV. NUMERICAL RESULTS

A CF-mMIMO network is considered where $L = 100$ APs, each equipped with $N = 4$ antennas, are uniformly distributed at random within a square coverage area of side 1000 m. To approximate a coverage area without boundaries, the nominal area is wrapped-around by eight identical neighbor replicas. Except otherwise stated, the same simulation parameters used in [33, Table V.4] have been adopted. Specifically, the path-losses have been fixed as $-30.5 - 36.7\log_{10}(d) + F$ (measured in dB), where d is the 3D distance (measured in meters) between the user and the AP, and F is the shadow fading, whose standard deviation and spatial decorrelation distance are set to 4 dB and 9 meters, respectively. The size of the TDD frame has been set to $\tau_{\mathrm{c}} = 200$ samples and a maximum pilot length of $\tau_{\mathrm{p}} = 20$ samples has been considered (i.e., the minimum DL payload data transmission size is equal to $\tau_{\mathrm{d}} = 180$ samples). The transmit power per pilot symbol is set to $P_{\mathrm{p}} = 100$ mW, whereas the available transmit power at the APs is set to $P_{\mathrm{dl}} = 200$ mW. The *virtual* UL transmit power used to generate the centralized IP-MMSE precoder is set to $p_g = 100$ mW. The parameters governing the DL power control policy have been set targeting max-min fairness solution, that is, $\nu = -0.5$ and $\kappa = 0.5$ for centralized IP-MMSE precoding [33, Fig. 7.2], and $\nu = 0.5$ for distributed CB precoding [33, Fig. 7.3]. An angular standard deviation of 15° has been considered when deriving the spatial correlation matrices using a local scattering model. Finally, a noise power spectral density of −174 dBm/Hz, a receiver noise figure of 7 dB, and an operating bandwidth of 20 MHz have been considered. Remarkably, each simulation result has been obtained as the average of 250 different snapshots of randomly deployed users and APs, with 500 fast-fading channel realizations per snapshot. Since all users in a given multicast subgroup transmit at the same rate (that of the worst user), the total sum SE metric helps to better visualize the effect subgrouping has on the network performance.

Figure 2 presents examples of simulation scenarios featuring different user distributions. Figure 2a illustrates a uniform distribution of 100 APs and 500 multicast users in a square area of side 1000 m. Figure 2b presents a uniform



(a) Uniform distribution of the multicast users (b) Clustered distribution of the multicast users (c) Heterogeneous distribution of the multicast users

FIGURE 2. Different distributions of 500 multicast users (black squares) in a square area of side 1 km served by 100 APs (red triangles) uniformly distributed at random. In the clustered distribution, the 500 multicast users are randomly dropped in 10 square clusters of side 50 m (50 users per cluster). In the heterogeneous distribution, 20 users are uniformly distributed in the coverage area, while the remaining users are grouped in clusters with different size: 2 clusters with 10 users, 3 clusters with 20 users, 5 clusters with 30 users, and 5 clusters with 50.

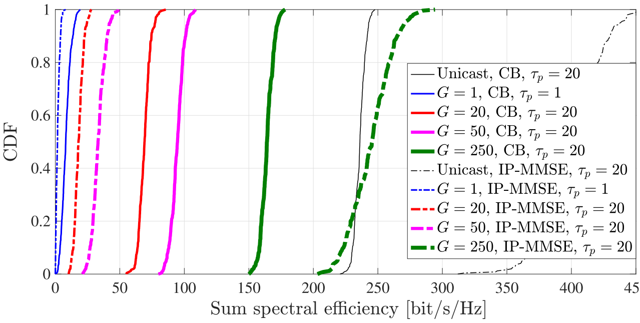


FIGURE 3. CDF of the sum SE. Uniform distribution of 500 multicast users. Unicast vs multicast with CB and IP-MMSE precoding. $L = 100$ APs, $N = 4$.

distribution of 100 APs with a clustered distribution of the multicast users in 10 square randomly distributed clusters of side 50 m, with 50 multicast users placed in each cluster. Figure 2c displays a uniform distribution of 100 APs with a heterogeneous distribution of the multicast users, uniformly and moderately clustered distributed.

A. UNIFORM DISTRIBUTION OF MULTICAST USERS

Figure 3 illustrates the cumulative distribution function (CDF) of the sum SE achieved by 500 uniformly distributed multicast users within the coverage area. We evaluate the performance of either unicast, single multicast, or multicast subgrouping transmissions.

The results show that irrespective of the precoding scheme, increasing the number of multicast subgroups improves the performance. As Figure 3 illustrates, creating 500 multicast subgroups with a single user per subgroup, that is, using unicast transmissions, IP-MMSE largely outperforms CB precoding as it occurs when conventional unicast transmissions are considered [33]. Using also $G = 250$ multicast subgroups, that is, each subgroup has an average of 2 multicast users, IP-MMSE presents significantly better sum SE results than the CB precoder. As we can observe, when the multicast users are uniformly distributed, IP-MMSE with unicast transmissions largely outperforms any multicast subgrouping option, from $G = 1$ single multicast group to $G = 250$ multicast subgroups.

It is worth highlighting that when the multicast users are uniformly distributed, the use of unicast transmission significantly outperforms any option employing multicast transmission, under both centralized IP-MMSE and distributed CB precoding. The superiority of unicast stems from the reasonable accuracy of the channel estimation and resulting precoder despite the pilot contamination (note that only $\tau_{\mathrm{p}} = 20$ orthogonal pilots are available in a scenario populated with $K = 500$ users). In other words, multicast transmissions are of little use in scenarios where the multicast users are uniformly distributed in a wide coverage area. Interestingly, in real-world scenarios, users are not likely to be uniformly distributed.

Figure 3 shows another interesting result. Analyzing the sum SE achieved using subgroup multicasting, we observe that CB outperforms IP-MMSE precoding as the number of users per subgroup increases. In other words, reducing the number of multicast subgroups improves the performance advantage of CB precoding (note that with $G = 1$, there are 500 users in a single multicast subgroup, whereas with $G = 50$, there are only an average of 10 users per multicast subgroup). Increasing the number of users per multicast subgroup causes significant intra-subgroup pilot contamination, which greatly deteriorates the quality of the channel estimates for each user in the subgroup. That is, although the users forming a specific multicast subgroup have similar channel statistics, the composite channel estimation for these users does not accurately reflect the propagation channel experienced by each individual user in the subgroup. Since IP-MMSE precoding is more sensitive to the degradation of channel estimates compared to the CB scheme, the SE provided by the IP-MMSE-based strategy suffers greater degradation than that provided by the CB scheme as the number of multicast subgroups decreases. In other words, any MMSE-based precoding scheme would significantly outperform CB precoding if accurate channel estimations were available. However, the use of a composite channel estimation process leads to inaccurate individual channel estimates as the number of users in each multicast subgroup increases. Unfortunately, the IP- MMSE strategy is much more sensitive to channel estimation inaccuracies than the CB scheme.

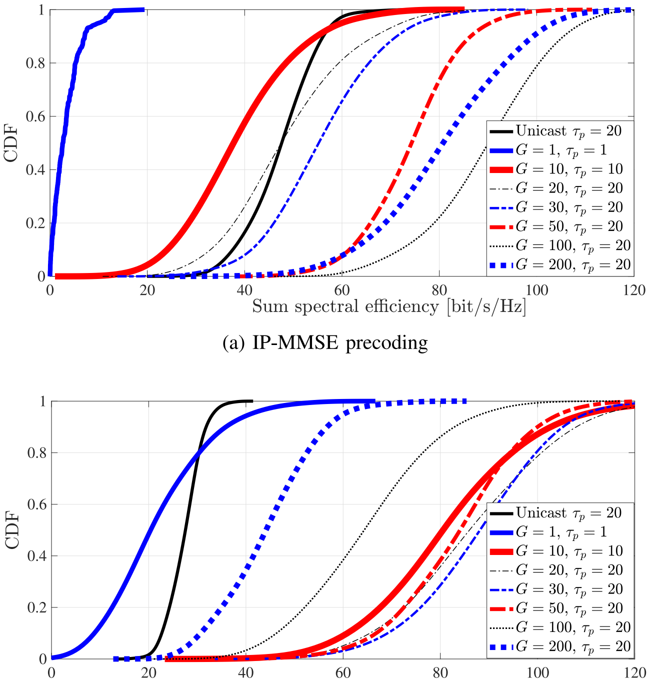
Measurable gains: In summary, multicast transmissions in scenarios with uniform distribution of users degrade the sum SE with respect to using unicast transmissions. IP- MMSE unicasting provides 1.6× more 95%-likely sum SE than CB unicast precoding, 5.8× more than $G = 20$ subgroups CB precoding, and 133× more than $G = 1$ single multicast group CB precoding.

B. CLUSTERED DISTRIBUTION OF MULTICAST USERS

To validate the utilization of multicast transmissions, and in contrast to the uniform user distribution just examined, we now deploy scenarios where the multicast users are located in square cluster areas of side 10 m, thus resulting in groups of users located very close to one another (e.g., lecture theaters, stations). This situation can extremely affect the channel estimation and the precoding due to pilot contamination. Fig. 4 shows the CDF of the sum SE of the multicast service when the $K = 500$ users are placed in 10 spatial clusters of 50 users each when using unicast, single multicast, or multicast subgrouping transmissions. Furthermore, we employ both centralized IP-MMSE and distributed CB precoding to deliver the service to all the users.

1) CENTRALIZED IP-MMSE PRECODING

Fig. 4(a) shows that centralized IP-MMSE precoding achieves the highest sum SE when transmitting to $G = 100$ groups. Remarkably, this strategy tends to approach unicast transmission (i.e., a larger number of subgroups, fewer users per subgroup) while preventing pilot contamination among inter-subgroup users located in the same spatial cluster (i.e., 100 multicast subgroups leads to an average of 10 subgroups per spatial cluster and $\tau_p = 20$). IP-MMSE allows the system to mitigate interference from closely located subgroups. The sole reason for not employing unicast is the significant pilot contamination from neighboring users. Given that the use of $\tau_p = 20$ samples provides only 20 orthogonal sequences, inter-subgroup pilot contamination is unavoidable in a unicast scenario with $G = 50$ multicast subgroups, each containing a single user located in a densely populated area. This contamination particularly affects the performance of IP-MMSE precoding. Indeed, it should be noted that the predominant impact of inter-subgroup pilot contamination emerges even when employing $G = 200$ subgroups, which degrades compared to the case in which $G = 100$ subgroups.



(a) IP-MMSE precoding

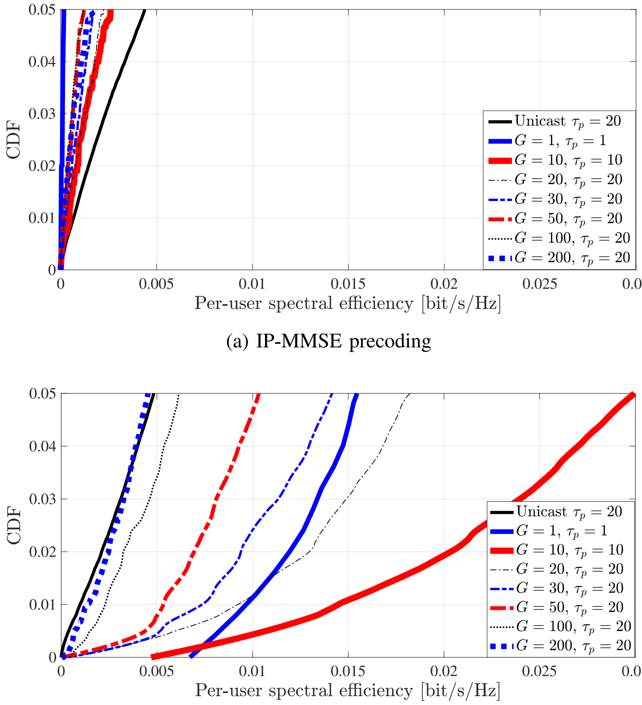
(b) CB precoding

FIGURE 4. CDF of the sum SE. Clustered distribution of multicast users with 10 spatial clusters (square cluster areas of side 10 m), of 50 users. Unicast vs multicast with IP-MMSE and CB precoding. L = 100 APs, N = 4.

2) DISTRIBUTED CB PRECODING

In contrast, Fig. 4(b) reveals that when using CB precoding, splitting the users into far fewer subgroups ($G = 30$ multicast subgroups) exhibits the best trade-off between useful signal and interference. This behavior can be understood by noting that while increasing the number of subgroups leads to better channel estimates (i.e., few users per multicast subgroup), going beyond a certain level of partitioning causes excessive cross-interference among subgroups located in the same spatial cluster. That is, CB precoding does not manage inter-subgroup interference and this is a factor in enhancing the CB performance, so the fewer number of multicast subgroups in a densely populated spatial area the lower the inter-subgroup interference. Opposite to IP-MMSE precoding, not only does the inter-subgroup pilot contamination affect the CB performance, but also the inter-subgroup DL interference. We should remark that the unicast transmissions are severely degraded due to the strong pilot contamination among users located in the same spatial cluster (note that 50 users are sharing $\tau_\text{p} = 20$ orthogonal pilot sequences). Additionally, it can be inferred from Fig. 4(b) that using a single multicast transmission does not yield the best sum SE performance. Therefore, it can be concluded that creating multicast subgroups based on users' locations outperforms both unicast and single multicast transmissions in scenarios where multicast users are distributed in spatial clusters.

Measurable gains: In summary, multicast subgrouping is the optimal strategy when the users are distributed in spatial



(a) IP-MMSE precoding

(b) CB precoding

FIGURE 5. **CDF of the per-user SE. Clustered distribution of multicast users with 10 spatial clusters (square cluster areas of side 10 m), of 50 users. Unicast vs multicast with IP-MMSE and CB precoding. $L = 100$ APs, $N = 4$.**

clusters. $G = 100$ subgroups IP-MMSE precoding provides 1.9× more 95%-likely sum SE than IP-MMSE unicast precoding, and 720× more than $G = 1$ single multicast group IP-MMSE precoding. Whereas $G = 30$ subgroups CB precoding provides 3× more 95%-likely sum SE than CB unicast precoding, and 10× more than $G = 1$ single multicast group CB precoding.

C. EVALUATION OF THE PER-USER SPECTRAL EFFICIENCY

So far, results have been shown for the sum SE to reveal the overall performance difference among different precoders and subgrouping strategies. Nonetheless, it is also important to re-examine these techniques on the basis of per-user performance, as the power allocation policy implemented supports both fairness and sum rate for the centralized power control, and slightly promotes the sum rate, for the distributed power control. In this subsection, we assess the performance of both centralized IP-MMSE and distributed CB precoding when using unicast, single multicast, and multicast subgrouping to evaluate the per-user SE. Figure 5 illustrates these results for the 95%-likely SE per multicast user.

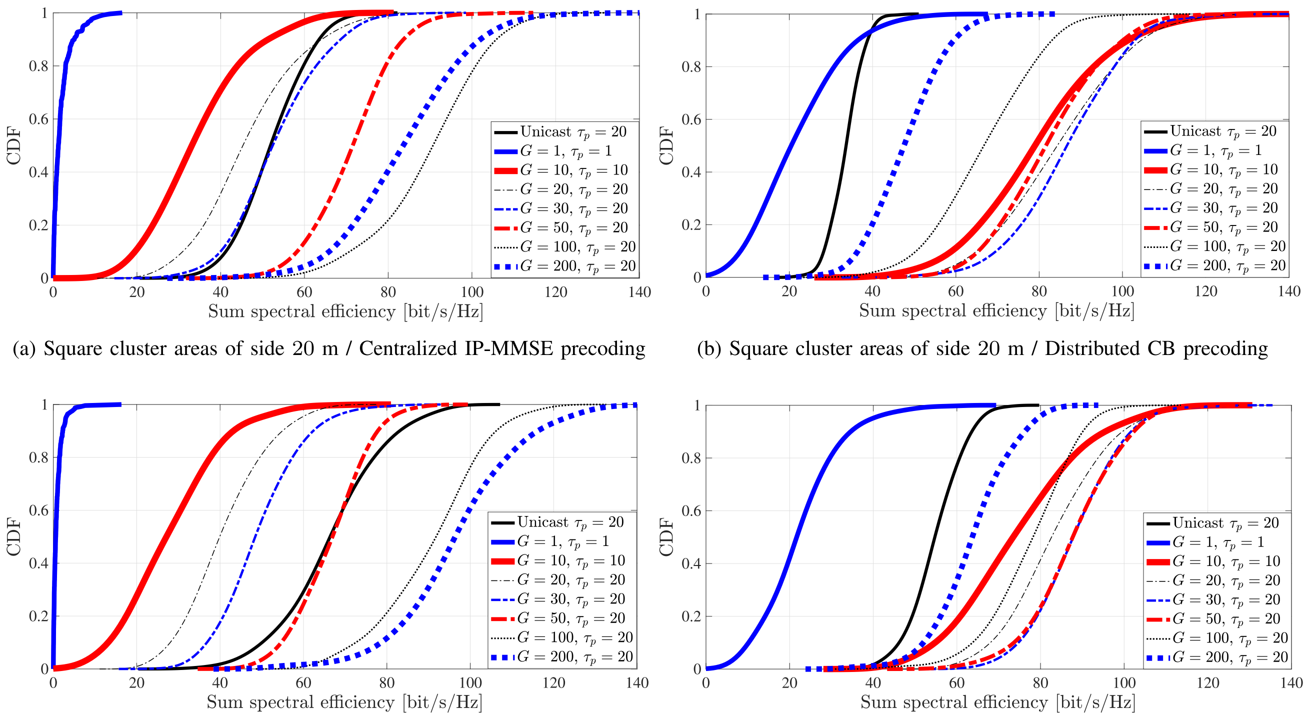
1) CENTRALIZED IP-MMSE PRECODING

Fig. 5(a) shows the results obtained when employing the centralized IP-MMSE precoder. We observe that the maximum 95%-likely per-user SE is obtained by unicast transmissions achieving almost 0.005 bit/s/Hz. Analyzing the sum SE and the per-user SE with the IP-MMSE precoding (see Figs. 4(a) and 5(a)), we can conclude that the configuration that provides the highest sum SE consists of creating 100 multicast subgroups, achieving approximately a 95%-likely sum SE of 67.6 bit/s/Hz. However, we observe that this configuration provides only a 95%-likely per-user SE of 0.0036 bit/s/Hz. Note that the configuration providing the highest 95%-likely per-user SE consists of unicast transmissions, delivering almost 0.005 bit/s/Hz, but achieving a 95%-likely sum SE of only 35.5 bit/s/Hz.

2) DISTRIBUTED CB PRECODING

On the other hand, Fig. 5(b) presents the per-user SE using the distributed CB precoding. The best option for 95%-likely per-user SE consists of creating 10 multicast subgroups achieving close to 0.03 bit/s/Hz, thus more than six times that achieved under IP-MMSE. The sum SE and the per-user SE with the CB precoding (Figure 4(b)and 5(b)) reflect that creating a moderate number of subgroups (i.e., 10) becomes a good compromise between both results. Although creating 30 subgroups allows the system to achieve a 95%-likely sum SE of approximately 65.3 bit/s/Hz, the use of only 10 multicast subgroups can achieve a 95%-likely sum SE of 57.3 bit/s/Hz. Nevertheless, the 95%-likely per-user SE obtained is close to 0.03 for 10 subgroups, doubling the 95%-likely per-user SE for 30 subgroups. It is interesting to observe how the distributed CB precoding notably outperforms, in terms of 95%-likely per-user SE, the centralized IP-MMSE precoding. We should notice that IP-MMSE achieves the best sum SE results using unicast transmissions. However, simulation results in the clustered scenarios show how some users' locations are significantly better than others, degrading the 95%-likely per-user SE. This situation is not improved using IP-MMSE with multicast subgroups since the intra-subgroup pilot contamination provides an inaccurate individual channel estimate that deteriorates the IP-MMSE precoder performance. In fact, grouping closely-located users to employ the composite channel estimate results in lower degradation in CB precoder performance. In addition, this grouping strategy minimizes the DL interference since the subgroups will be not closely located. Note that all the multicast users belonging to the same subgroup will benefit from the same per-user SE. Thus, CB precoding demonstrates superior performance as the number of users per multicast subgroup increases, as it is less sensitive to channel estimation inaccuracies than IP-MMSE, ultimately resulting in a higher 95%-likely per-user SE.

After the performance analysis of centralized and distributed precoding strategies in the clustered distribution of multicast users both with sum SE and per-user SE, it should be noted that the highest sum SE achieved with the IP-MMSE option is very similar to that obtained using the best CB option. In such cases, other factors such as the per-user SE and the computational complexity of the precoder might be relevant to consider. The per-user SE results



(a) Square cluster areas of side 20 m / Centralized IP-MMSE precoding

(b) Square cluster areas of side 20 m / Distributed CB precoding

(c) Square cluster areas of side 50 m / Centralized IP-MMSE precoding

(d) Square cluster areas of side 50 m / Distributed CB precoding

FIGURE 6. CDF of the sum SE. Clustered distribution of multicast users with 10 spatial clusters of 50 users. Different sizes of the cluster areas. Unicast vs multicast with IP-MMSE and CB precoding. $L = 100$ APs, $N = 4$.

push us to recommend the use of distributed CB precoding with multicast subgrouping in clustered scenarios since it allows the majority of the users to achieve a higher SE performance. Moreover, note that in a similar manner to the power allocation, which can be set following a prescribed rate policy (e.g., sumrate, maxmin), so does the number of subgroups G, which can also be set targeting different objectives. These objectives are hard to find analytically in practice making the optimization problem unaffordable with conventional optimization techniques. Deep learning techniques could be an appropriate way to determine the optimal number of multicast subgroups depending on the network requirements, precoding scheme, and users' and APs' distribution. Note that the optimization problem depends on large-scale parameters and can be optimized whenever the spatial distribution changes significantly. Furthermore, the solution to the optimization problem depends on the variable to optimize, that is, sum SE, per-user SE, or EE. The computational complexity will be an important factor in deciding the precoding strategy. We should emphasize that the lower the number of multicast subgroups, the lower the computational complexity. Thus, to minimize the computational complexity we recommend employing CB precoding that offers a lower computational complexity and provides a better performance for multicast subgroups with more users.

Measurable gains: In summary, multicast subgrouping with CB precoding is the optimal strategy when the users are distributed in spatial clusters to maximize the 95%-likely per-user SE. $G = 10$ subgroups CB precoding provides 1.9× more 95%-likely per-user SE than $G = 1$ single multicast group CB precoding, 6.2× more than CB unicast precoding, and 6.8× more than IP-MMSE unicast precoding.

D. EFFECT OF THE SPATIAL CLUSTER AREA

In this subsection, we study the effect of the size of the cluster areas where the multicast users are distributed. In Fig. 6, results are presented for the sum SE when square cluster areas of side 20 m and 50 m are employed.

1) CENTRALIZED IP-MMSE PRECODING

Figure 6(a) and 6(c) illustrate the sum SE using centralized IP-MMSE precoding with square cluster areas of side 20 m and 50 m, respectively. Comparing these results with those presented in Fig. 4(a) (square clusters with a side length of 10 m), we observe similar trends, with neither unicast nor a single multicast emerging as the best option. Nonetheless, it is interesting to note that the larger the area of the spatial cluster of multicast users, the larger the number of multicast subgroups leading to the optimal sum SE.

In particular, whereas using $G = 100$ subgroups is the best-evaluated option when users are distributed in square cluster areas of side 20 m, using $G = 200$ subgroups is the best option when increasing the square cluster area side to 50 m. Note that the larger the spatial cluster area where the users are located is, the lower the density of users in this area is, and consequently, the closer the scenario becomes to a uniform distribution of users. While it is true that distributing users in square spatial clusters with sides of 50 m results in a distribution that is far from uniform, there is a tendency towards a more uniform distribution, which favors the IP-MMSE strategy.

Results for the extreme case have been depicted in Figure 3, wherein multicast users are uniformly distributed within a square area of side 1000 meters. In this scenario, unicast streams (i.e., $G = 500$ subgroups) are the preferred transmission option for delivering the multicast service.

2) DISTRIBUTED CB PRECODING

In Figure 6(b) and 6(d), the sum SE results obtained using distributed CB precoding for square cluster areas with side lengths of 20 m and 50 m are presented. We observe similar trends to those already presented in Fig. 4b (square clusters with a side length of 10 m), observing that CB precoding shows better performance with moderate G size ($G = 20$, $G = 30$, $G = 50$) in contrast to IP-MMSE precoding ($G = 100$, $G = 200$).

Furthermore, we should notice that increasing the cluster area size allows both the IP-MMSE and CB precoders to achieve higher sum SEs as the number of subgroups increases. Specifically, a larger optimal number of multicast subgroups is obtained as the spatial cluster area increases. Nevertheless, as depicted in Figure 4(b), when striving for the highest SE, employing distributed CB precoding leads to fewer multicast subgroups (each with a larger number of users) compared to centralized IP-MMSE precoding.

It is worth emphasizing that regardless of the cluster area, whenever users are distributed within those spatial clusters, creating multicast subgroups based on the spatial locations of the users leads to better sum SE results compared to those achieved under unicast or single multicast transmission.

Measurable gains: In summary, multicast subgrouping continues appearing as the optimal strategy when the users are distributed in larger spatial clusters (50 m side). $G = 200$ subgroups IP-MMSE precoding provides 2.4× more 95%-likely sum SE than IP-MMSE unicast precoding, and 5.5× more than single $G = 1$ multicast group IP-MMSE precoding. While $G = 30$ subgroups CB precoding provides 1.6× more 95%-likely sum SE than CB unicast precoding, and 9.4× more than single $G = 1$ multicast group CB precoding.

E. EFFECT OF THE SPATIAL CORRELATION AT THE APS

This subsection provides insights into how the spatial correlation among the antennas of the APs impacts the performance of multicast strategies. Fig. 7 illustrates the sum SE for highly correlated antennas (covariance matrices obtained using azimuth and elevation angular standard deviations (ASDs) $\sigma_\varphi = \sigma_\theta = 5°$), and for uncorrelated Rayleigh fading channels. The multicast users are assumed to be located in square cluster areas of side 10 m.

1) CENTRALIZED IP-MMSE PRECODING

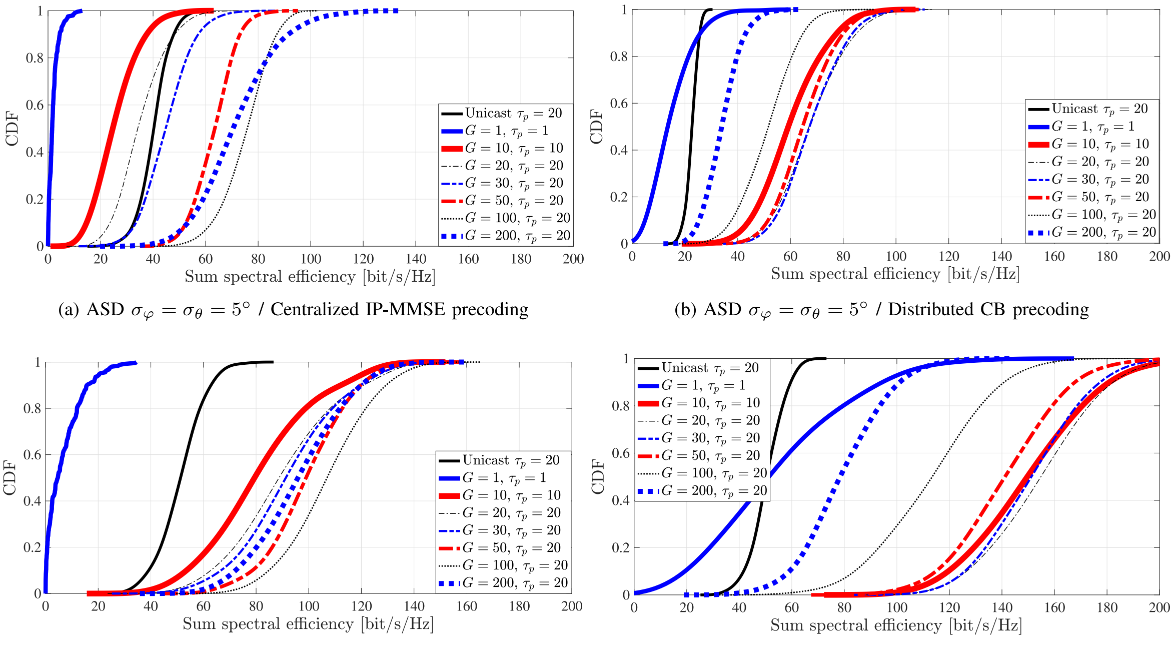
Analyzing the results in Figure 7(a) and 7(c), obtained using centralized IP-MMSE precoding, it becomes apparent that spatial correlation deteriorates performance, particularly evident with fewer multicast subgroups and consequently, with larger numbers of users per subgroup because spatial correlation degrades the composite channel estimation and, consequently, the precoder design is less accurate than in the uncorrelated Rayleigh fading scenario. This is in stark contrast to the situation in unicast transmissions, where spatial correlation is known to be advantageous when users are uniformly distributed (as reported in [33]), and the channel estimation is carried out individually. In this latter case (i.e., in a clustered scenario), note that the benefits spatial correlation brings are somewhat overshadowed by the impact of pilot contamination since the number of users located in a spatial cluster area (50 users per subgroup) is higher than the number of available orthogonal pilots ($\tau_p = 20$ samples), thus the correlation matrices characterizing the channels experienced by those users are very similar and the pilot contamination cannot be compensated under spatially correlated fading channels. While results remain largely unaffected by spatial correlation when implementing the $G = 200$ subgroup option (i.e., with low levels of intra-subgroup pilot contamination), implementing options with fewer multicast subgroups formed by a larger number of users per subgroup (i.e., with high levels of intra-subgroup pilot contamination) leads to a deterioration in the sum SE. This decline occurs because the quality of common channel estimation and the effectiveness of precoder design degrade due to spatial correlation among antennas at the APs. This trend can be easily observed in Figure 7(a) and 7(c) when looking at the 1- or 10-subgroup options.

2) DISTRIBUTED CB PRECODING

Figure 7(b) and 7(d) reflect the sum SE results using distributed CB precoding in scenarios with highly correlated antennas at the APs and scenarios with uncorrelated Rayleigh fading channels, respectively. It can be observed that the deleterious effects of spatial correlation are notably higher not only when creating a few subgroups with a large number of users (e.g., 1- or 10-subgroup options) but also when using unicast or a large number of subgroups with a small number of users per subgroup (e.g., 100- or 200-subgroup options).

The deteriorating performance of distributed CB precoding in the presence of spatial correlation, as highlighted in [33], is further exacerbated in scenarios where users are spatially clustered, primarily due to the adverse impacts of pilot contamination.

Furthermore, spatial correlation, as occurs with centralized IP-MMSE, degrades the common channel estimation and the subsequent precoder design when compared to the uncorrelated Rayleigh fading scenario. As it has been pointed out before, irrespective of the degree of spatial correlation, multicast subgrouping exhibits important enhancements in the achieved sum SE with respect to the unicast or the single-group multicast schemes.



(a) ASD $\sigma_\varphi = \sigma_\theta = 5^\circ$ / Centralized IP-MMSE precoding

(b) ASD $\sigma_\varphi = \sigma_\theta = 5^\circ$ / Distributed CB precoding

(c) Uncorrelated Rayleigh fading / Centralized IP-MMSE precoding

(d) Uncorrelated Rayleigh fading / Distributed CB precoding

FIGURE 7. CDF of the sum SE. Clustered distribution of multicast users with 10 spatial clusters square (cluster areas of side 10 m) of 50 users. Different spatially correlated Rayleigh fading ASDs. Unicast vs multicast with IP-MMSE and CB precoding. $L = 100$ APs, $N = 4$.

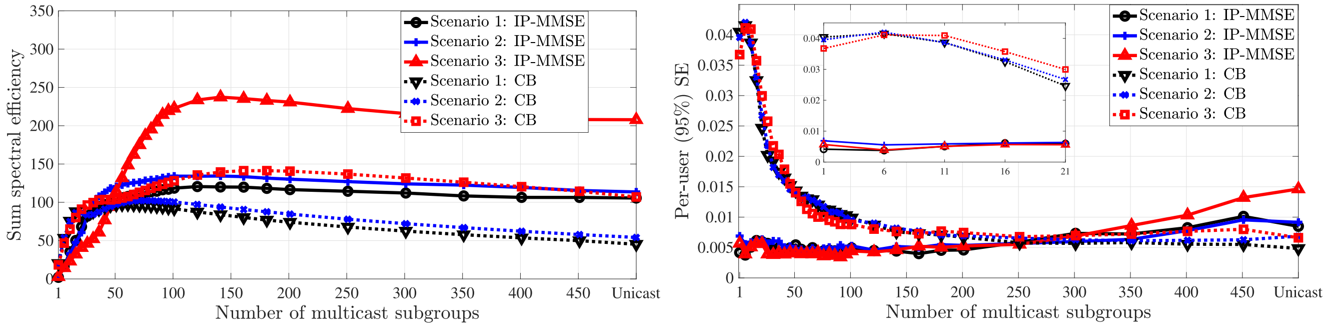
Measurable gains: In summary, the spatial correlation at the APs antennas degrades the sum SE when the multicast users are distributed in spatial clusters. This degradation is significantly larger using CB than IP-MMSE precoder. $G = 100$ subgroups IP-MMSE precoding with uncorrelated Rayleigh fading channels provides 1.4× more 95%-likely sum SE than the same precoding with high correlated Rayleigh fading channels. While $G = 20$ subgroups CB precoding with uncorrelated Rayleigh fading channels provides 2.5× more 95%-likely sum SE than the same precoding with high correlated Rayleigh fading channels.

F. ANALYSIS OF HETEROGENEOUS DISTRIBUTION OF MULTICAST USERS

Finally, this subsection explores the impact different heterogeneous distributions of multicast users have on spectral efficiency. These distributions present different mixes in the numbers of uniformly distributed users, spatial clusters, and users per spatial cluster that allow us to analyze the effect of multicast subgrouping in a heterogeneous setting. We consider a fixed number of $K = 500$ users, populating a square area of side 1000 m under three different distributions of users:

- *Scenario 1 (highly clustered)*: 10 users uniformly distributed, 2 spatial clusters of 10 users, 4 spatial clusters of 20 users, 3 spatial clusters of 30 users, and 6 spatial clusters of 50 users.
- *Scenario 2 (moderately clustered)*: 20 users uniformly distributed, 2 spatial clusters of 10 users, 3 spatial clusters of 20 users, 5 spatial clusters of 30 users, and 5 spatial clusters of 50 users.
- *Scenario 3 (sparsely clustered)*: 100 users uniformly distributed, 5 spatial clusters of 10 users, 5 spatial clusters of 20 users, 5 spatial clusters of 30 users, and 2 spatial clusters of 50 users.

Figure 8 presents the sum SE and the 95%-likely per-user SE in the described scenarios using both centralized IP-MMSE and distributed CB precoding. Regarding the sum SE results, it can be confirmed how neither using unicast nor single-group multicast results in the highest sum SE in any of the assessed spatial distributions of users and precoding strategies. The number of multicast subgroups that maximizes the sum SE depends on the distribution of the multicast users. Clearly, by increasing the number of uniformly distributed users, transitioning from a highly clustered scenario to a sparsely clustered one, the number of multicast subgroups needed to maximize the sum SE also increases. Note that most of these subgroups consist of only one user (unicast transmission), in fact matching the number of uniformly distributed users. It is also noticeable that the sum SE achieved using centralized IP-MMSE precoding outperforms those obtained using distributed CB precoding. The higher the number of uniformly distributed users, the larger the benefit of using IP-MMSE in terms of sum SE. This result can be explained by the enormous benefit in unicast transmissions brought by IP-MMSE in comparison to what is achieved under CB precoding. When increasing the number of multicast subgroups (i.e., increasing the number of unicast transmissions) beyond its optimal operational point, it is found that IP-MMSE exhibits a more gradual degradation in terms of SE than CB. It can be concluded that centralized IP-MMSE precoding employs as many unicast transmissions



(a) Sum SE

(b) 95%-likely per-user SE

FIGURE 8. Sum SE and 95%-likely per-user SE vs number of multicast subgroups. 500 multicast users with different heterogeneous distributions. Centralized IP-MMSE vs distributed CB precoding. $L = 100$ APs, $N = 4$.

as strong pilot contamination permits. In contrast, distributed CB precoding maximizes the sum SE by grouping users based on their large-scale channel similarities.

Turning now our attention to Figure 8(b), the 95%-likely per-user SE shows that creating a small number of multicast subgroups (i.e., around 10 subgroups depending on the users' distribution) when relying on the distributed CB precoding outperforms any other option (see zoomed region). The per-user SE using CB precoding shows that multicast subgrouping outperforms unicast and single multicast options. In contrast, using IP-MMSE precoding, the 95%-likely per-user SE increases with the number of unicast transmissions, ultimately making unicast transmission virtually optimal.

It is worth emphasizing once more that even in heterogeneous user deployments, multicast subgrouping outperforms the sum SE results achieved through unicast and single multicast transmissions.

V. CONCLUSION

This work has considered CF-mMIMO multicasting under spatially correlated Rayleigh fading channels. In particular, a novel framework has been proposed to assess the performance of scalable multicast techniques in CF-mMIMO networks when using different precoders and power allocation strategies. Moreover, a novel subgrouping technique has been proposed whereby multicast users are separated based on their spatial location aiming at improving the performance of the multicast service when users are not uniformly distributed.

An exhaustive numerical evaluation reveals the benefits brought by multicast subgrouping when users tend to form spatial clusters defining hotspots with larger user densities than those found in uniform distributions. The main conclusions derived from this work are:

- Unicast transmissions utilizing centralized IP-MMSE precoding are preferred when multicast users are uniformly distributed. However, when multicast users tend to form spatial clusters, unicast transmissions suffer severe degradation due to pilot contamination, whereas multicast subgrouping can maintain significantly higher rates. The advantages of multicast subgrouping persist even with the expansion of the areas over which users are spatially distributed (i.e., less densely populated spatial clusters) or variations in spatial correlation among the arrays of antennas at the APs.
- Distributed CB precoding improves centralized IP-MMSE sum SE when a moderate number of multicast subgroups is employed. Furthermore, distributed CB precoding achieves significantly better results in terms of 95%-likely per-user SE than centralized IP-MMSE precoding.
- The heterogeneous distributions of users, encompassing clusters with varying population densities alongside users uniformly distributed across the network coverage area, confirm that multicast subgrouping yields the highest sum SE when employing IP-MMSE precoding. Additionally, it demonstrates that CB precoding with a small number of subgroups achieves the best 95%-likely per-user SE.

It has been demonstrated that grouping users based on their spatial channel similarities notably enhances the efficacy of multicast services in scenarios featuring a dense concentration of nearby users. An intriguing avenue for future research would be exploring deep learning techniques to determine the optimal number of multicast subgroups. This determination is linked to the average channel gain vectors of users in the system, which, in turn, are somewhat associated with their respective locations. This suggests that leveraging a priori knowledge of users' average channel gain vectors could enhance the accuracy of finding the optimal number of multicast subgroups. Additionally, leveraging the derived SE closed-form expressions for CB, it appears plausible to devise novel power allocation strategies aiming at maximizing either the sum SE or the minimum per-user SE.

APPENDIX A
PROOF OF LEMMA 1

The MMSE estimate of the subgroup channel $\boldsymbol{h}_l^g$ can be obtained as

$$\begin{aligned}\hat{\boldsymbol{h}}_l^g &= \mathbb{E}\{\boldsymbol{h}_l^g | \boldsymbol{y}_l^g\}\\ &= \mathbb{E}\left\{\boldsymbol{h}_l^g (\boldsymbol{y}_l^g)^{\mathsf{H}}\right\}\left(\mathbb{E}\left\{\boldsymbol{y}_l^g (\boldsymbol{y}_l^g)^{\mathsf{H}}\right\}\right)^{-1} \boldsymbol{y}_l^g,\end{aligned} \tag{29}$$

where

$$\mathbb{E}\left\{\boldsymbol{h}_l^g\left(\boldsymbol{y}_l^g\right)^{\mathsf{H}}\right\} = K_g\mathbb{E}\left\{\boldsymbol{h}_{lk}\left(\boldsymbol{h}_{lk}\right)^{\mathsf{H}}\right\} = K_g\boldsymbol{R}_l^g, \tag{30}$$

and

$$\begin{aligned}\mathbb{E}&\left\{\boldsymbol{y}_l^g\left(\boldsymbol{y}_l^g\right)^{\mathsf{H}}\right\}\\ &= \tau_{\mathrm{p}}P_{\mathrm{p}}\sum_{\substack{c\in\mathcal{G}\\ \boldsymbol{\psi}_c=\boldsymbol{\psi}_g}}\sum_{i\in\mathcal{K}_c}\mathbb{E}\left\{\boldsymbol{h}_{li}\boldsymbol{h}_{li}^{\mathsf{H}}\right\} + \mathbb{E}\left\{\boldsymbol{n}_{lg}\boldsymbol{n}_{lg}^{\mathsf{H}}\right\}\\ &= \tau_{\mathrm{p}}P_{\mathrm{p}}\sum_{\substack{c\in\mathcal{G}\\ \boldsymbol{\psi}_c=\boldsymbol{\psi}_g}}\sum_{i\in\mathcal{K}_c}\boldsymbol{R}_{li} + \sigma_{\mathrm{u}}^2\boldsymbol{I}_N \triangleq \boldsymbol{\Gamma}_{lg}. \end{aligned}\tag{31}$$

The MMSE estimate is a zero-mean complex Gaussian vector whose correlation matrix can be straightforwardly obtained as

$$\mathbb{E}\left\{\hat{\boldsymbol{h}}_l^g\left(\hat{\boldsymbol{h}}_l^g\right)^{\mathsf{H}}\right\} = K_g^2\boldsymbol{R}_l^g\boldsymbol{\Gamma}_{lg}^{-1}\boldsymbol{R}_l^g. \tag{32}$$

The subgroup channel estimation error $\tilde{\boldsymbol{h}}_l^g = \boldsymbol{h}_l^g - \hat{\boldsymbol{h}}_l^g$ is also a zero-mean complex Gaussian vector whose correlation matrix is given by

$$\begin{aligned}\mathbb{E}\left\{\tilde{\boldsymbol{h}}_l^g\left(\tilde{\boldsymbol{h}}_l^g\right)^{\mathsf{H}}\right\} &= \mathbb{E}\left\{\boldsymbol{h}_l^g\left(\boldsymbol{h}_l^g\right)^{\mathsf{H}}\right\} - \mathbb{E}\left\{\hat{\boldsymbol{h}}_l^g\left(\hat{\boldsymbol{h}}_l^g\right)^{\mathsf{H}}\right\}\\ &= \boldsymbol{R}_l^g - K_g^2\boldsymbol{R}_l^g\,\boldsymbol{\Gamma}_{lg}^{-1}\boldsymbol{R}_l^g. \end{aligned}\tag{33}$$

APPENDIX B PROOF OF LEMMA 2

The CB precoder can be expressed as

$$\mathbf{w}_{lg} = \frac{\sqrt{\rho_{lg}}\boldsymbol{D}_{lg}\hat{\boldsymbol{h}}_l^g}{\sqrt{\mathbb{E}\{\|\boldsymbol{D}_{lg}\hat{\boldsymbol{h}}_l^g\|^2\}}} = \frac{\sqrt{\rho_{lg}}\boldsymbol{D}_{lg}\hat{\boldsymbol{h}}_l^g}{K_g\sqrt{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^g\boldsymbol{R}_l^g\right)}} \tag{34}$$

where $\boldsymbol{\Lambda}_l^g = \boldsymbol{D}_{lg}\boldsymbol{R}_l^g\boldsymbol{\Gamma}_{lg}^{-1}$. Using (4) and (7) in this expression, and exploiting the uncorrelation between different channel vectors, it can be shown that

$$\begin{aligned}\mathbb{E}\{\varrho_{lk}^{gg}\} &= \sqrt{\frac{\rho_{lg}\tau_{\mathrm{p}}P_{\mathrm{p}}}{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^g\boldsymbol{R}_l^g\right)}}\,\mathbb{E}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^g\boldsymbol{h}_{lk}\right\}\\ &= \sqrt{\frac{\rho_{lg}\tau_{\mathrm{p}}P_{\mathrm{p}}}{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^g\boldsymbol{R}_l^g\right)}}\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^g\right). \end{aligned}\tag{35}$$

A general expression for the expectation in (27) can be obtained as

$$\begin{aligned}\mathbb{E}\left\{\left|\sum_{l=1}^{L}\varrho_{lk}^{gc}\right|^2\right\} &= \sum_{l=1}^{L}\sum_{l'=1}^{L}\mathbb{E}\left\{\varrho_{lk}^{gc}\left(\varrho_{l'k}^{gc}\right)^*\right\}\\ &= \sum_{l=1}^{L}\sum_{l'=1}^{L}\sqrt{\frac{\rho_{lc}\rho_{l'c}}{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^c\boldsymbol{R}_l^c\right)\mathrm{tr}\left(\boldsymbol{\Lambda}_{l'}^c\boldsymbol{R}_{l'}^c\right)}}\mathbb{E}\left\{\vartheta_{ll'k}^{gc}\right\}, \end{aligned}\tag{36}$$

where $\vartheta_{ll'k}^{gc} = \boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\boldsymbol{y}_l^c(\boldsymbol{y}_{l'}^c)^{\mathsf{H}}(\boldsymbol{\Lambda}_{l'}^c)^{\mathsf{H}}\boldsymbol{h}_{l'k},\ k\in K_g$. The computation of $\mathbb{E}\{\vartheta_{ll'k}^{gc}\}$ can be split in three different cases:

1) If $\boldsymbol{\psi}_c \neq \boldsymbol{\psi}_g$, then

$$\begin{aligned}\mathbb{E}\{\vartheta_{ll'k}^{gc}\} &= \mathbb{E}_{\boldsymbol{h},\boldsymbol{y}}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\boldsymbol{y}_l^c\left(\boldsymbol{y}_{l'}^c\right)^{\mathsf{H}}\left(\boldsymbol{\Lambda}_{l'}^c\right)^{\mathsf{H}}\boldsymbol{h}_{l'k}\right\}\\ &= \mathbb{E}_{\boldsymbol{h}}\left\{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{h}}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\boldsymbol{y}_l^c\left(\boldsymbol{y}_{l'}^c\right)^{\mathsf{H}}\left(\boldsymbol{\Lambda}_{l'}^c\right)^{\mathsf{H}}\boldsymbol{h}_{l'k}\middle|\boldsymbol{h}\right\}\right\}\\ &= \mathbb{E}_{\boldsymbol{h}}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\,\mathbb{E}_{\boldsymbol{y}}\left\{\boldsymbol{y}_l^c\left(\boldsymbol{y}_{l'}^c\right)^{\mathsf{H}}\right\}\left(\boldsymbol{\Lambda}_{l'}^c\right)^{\mathsf{H}}\boldsymbol{h}_{l'k}\right\}\\ &= \begin{cases}\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\boldsymbol{R}_l^c\boldsymbol{D}_{lc}\right), & \text{if}\ \ l' = l,\\ 0, & \text{if}\ l' \neq l.\end{cases} \end{aligned}\tag{37}$$

2) If $\boldsymbol{\psi}_c = \boldsymbol{\psi}_g$ and $l \neq l'$, then

$$\begin{aligned}\mathbb{E}\{\vartheta_{ll'k}^{gc}\} &= \mathbb{E}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\boldsymbol{y}_l^c\right\}\mathbb{E}\left\{\left(\boldsymbol{y}_{l'}^c\right)^{\mathsf{H}}\left(\boldsymbol{\Lambda}_{l'}^c\right)^{\mathsf{H}}\boldsymbol{h}_{l'k}\right\}\\ &= \tau_{\mathrm{p}}P_{\mathrm{p}}\,\mathrm{tr}\left(\boldsymbol{\Lambda}_l^c\boldsymbol{R}_{lk}\right)\mathrm{tr}\left(\left(\boldsymbol{\Lambda}_{l'}^c\right)^{\mathsf{H}}\boldsymbol{R}_{l'k}\right) \end{aligned}\tag{38}$$

3) If $\boldsymbol{\psi}_c = \boldsymbol{\psi}_g$ and $l = l'$, then

$$\begin{aligned}\mathbb{E}\{\vartheta_{ll'k}^{gc}\} &= \tau_{\mathrm{p}}P_{\mathrm{p}}\mathbb{E}\left\{\left|\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\boldsymbol{h}_{lk}\right|^2\right\}\\ &\quad + \tau_{\mathrm{p}}P_{\mathrm{p}}\sum_{k'\in\mathcal{K}_g\backslash\{k\}}\mathbb{E}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\,\mathbb{E}\left\{\boldsymbol{h}_{lk'}\boldsymbol{h}_{lk'}^{\mathsf{H}}\right\}\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\boldsymbol{h}_{lk}\right\}\\ &\quad + \tau_{\mathrm{p}}P_{\mathrm{p}}\sum_{\substack{c'\in\mathcal{G}\backslash\{g\}\\ \boldsymbol{\psi}_{c'}=\boldsymbol{\psi}_g}}\sum_{i\in K_{c'}}\mathbb{E}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\,\mathbb{E}\left\{\boldsymbol{h}_{li}\boldsymbol{h}_{li}^{\mathsf{H}}\right\}\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\boldsymbol{h}_{lk}\right\}\\ &\quad + \mathbb{E}_{\boldsymbol{h}}\left\{\boldsymbol{h}_{lk}^{\mathsf{H}}\boldsymbol{\Lambda}_l^c\,\mathbb{E}_{\boldsymbol{n}}\left\{\boldsymbol{n}_{lg}\boldsymbol{n}_{lg}^{\mathsf{H}}\right\}\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\boldsymbol{h}_{lk}\right\}\\ &= \tau_{\mathrm{p}}P_{\mathrm{p}}\left[\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\right)\right]^2 + \tau_{\mathrm{p}}P_{\mathrm{p}}\,\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\boldsymbol{R}_{lk}\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\right)\\ &\quad + \tau_{\mathrm{p}}P_{\mathrm{p}}\sum_{k'\in\mathcal{K}_g\backslash\{k\}}\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\boldsymbol{R}_{lk'}\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\right)\\ &\quad + \tau_{\mathrm{p}}P_{\mathrm{p}}\sum_{\substack{c'\in\mathcal{G}\backslash\{g\}\\ \boldsymbol{\psi}_{c'}=\boldsymbol{\psi}_g}}\sum_{i\in K_{c'}}\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\boldsymbol{R}_{li}\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\right)\\ &\quad + \sigma_{\mathrm{u}}^2\,\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\right)\\ &= \tau_{\mathrm{p}}P_{\mathrm{p}}\left[\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\right)\right]^2 + \mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{D}_{lc}\boldsymbol{R}_l^c\left(\boldsymbol{\Lambda}_l^c\right)^{\mathsf{H}}\right), \end{aligned}\tag{39}$$

where the last equality is obtained by using (8).

Finally, combining (36)–(39), we have

$$\begin{aligned}\mathbb{E}\left\{\left|\sum_{l=1}^{L}\varrho_{lk}^{gc}\right|^2\right\} &= \sum_{l=1}^{L}\rho_{lc}\frac{\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\boldsymbol{R}_l^c\boldsymbol{D}_{lc}\right)}{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^c\boldsymbol{R}_l^c\right)}\\ &\quad + \begin{cases}\left|\sum_{l=1}^{L}\sqrt{\rho_{lc}\tau_{\mathrm{p}}P_{\mathrm{p}}}\frac{\mathrm{tr}\left(\boldsymbol{R}_{lk}\boldsymbol{\Lambda}_l^c\right)}{\mathrm{tr}\left(\boldsymbol{\Lambda}_l^c\boldsymbol{R}_l^c\right)}\right|^2, & \text{if}\ \ \boldsymbol{\psi}_c = \boldsymbol{\psi}_g,\\ 0, & \text{if}\ \boldsymbol{\psi}_c \neq \boldsymbol{\psi}_g.\end{cases} \end{aligned}\tag{40}$$

REFERENCES

[1] "Ericsson mobility report," Ericsson, Stockholm, Sweden, White Paper, Nov. 2023, [Online] Available: https://www.ericsson.com/en/reports-and-papers/mobility-report

[2] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[3] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[4] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO Versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[5] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–13, 2019.

[6] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44596–44612, 2019.

[7] H. Q. Ngo, G. Interdonato, E. G. Larsson, G. Caire, and J. G. Andrews, "Ultradense cell-free massive MIMO for 6G: Technical overview and open questions," *Proc. IEEE*, early access, May 8, 2024, doi: 10.1109/JPROC.2024.3393514.

[8] K. Samdanis and T. Taleb, "The road beyond 5G: A vision and insight of the key technologies," *IEEE Netw.*, vol. 34, no. 2, pp. 135–141, Mar./Apr. 2020.

[9] X. You et al., "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China, Inf. Sci.*, vol. 64, pp. 1–74, Jan. 2021.

[10] M. Zhou, Y. Zhang, X. Qiao, and L. Yang, "Spatially correlated rayleigh fading for cell-free massive MIMO systems," *IEEE Access*, vol. 8, pp. 42154–42168, 2020.

[11] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[12] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[13] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[14] A. de la Fuente, R. P. Leal, and A. G. Armada, "New technologies and trends for next generation mobile broadcasting services," *IEEE Wireless Commun. Mag.*, vol. 54, no. 11, pp. 217–223, Nov. 2016.

[15] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Netw.*, vol. 31, no. 2, pp. 80–89, Mar./Apr. 2017.

[16] A. Papazafeiropoulos, H. Q. Ngo, P. Kourtessis, S. Chatzinotas, and J. M. Senior, "Towards optimal energy efficiency in cell-free massive MIMO systems," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 2, pp. 816–831, Jun. 2021.

[17] N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Approaching 6G use case requirements with multicasting," *IEEE Wireless Commun. Mag.*, vol. 61, no. 5, pp. 144–150, May 2023.

[18] V. K. Shrivastava, S. Baek, and Y. Baek, "5G evolution for multicast and broadcast services in 3GPP release 17," *IEEE Commun. Stand. Mag.*, vol. 6, no. 3, pp. 70–76, Sep. 2022.

[19] X. Lin, "An overview of 5G advanced evolution in 3GPP release 18," *IEEE Commun. Stand. Mag.*, vol. 6, no. 3, pp. 77–83, Sep. 2022.

[20] G. Araniti, M. Condoluci, L. Militano, and A. Iera, "Adaptive resource allocation to multicast services in LTE systems," *IEEE Trans. Broadcast.*, vol. 59, no. 4, pp. 658–664, Dec. 2013.

[21] A. de la Fuente, G. Femenias, F. Riera-Palou, and A. G. Armada, "Subband CQI feedback-based multicast resource allocation in MIMO-OFDMA networks," *IEEE Trans. Broadcast.*, vol. 64, no. 4, pp. 846–864, Dec. 2018.

[22] A. de la Fuente, G. Interdonato, and G. Araniti, "User subgrouping and power control for multicast massive MIMO over spatially correlated channels," *IEEE Trans. Broadcast.*, vol. 68, no. 4, pp. 834–847, Dec. 2022.

[23] H. Yang, T. L. Marzetta, and A. Ashikhmin, "Multicast performance of large-scale antenna systems," in *Proc. IEEE 14th Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2013, pp. 604–608.

[24] M. Sadeghi, L. Sanguinetti, R. Couillet, and C. Yuen, "Reducing the computational complexity of multicasting in large-scale antenna systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2963–2975, May 2017.

[25] M. Dong and Q. Wang, "Optimal multi-group multicast beamforming structure," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[26] T. X. Doan, H. Q. Ngo, T. Q. Duong, and K. Tourki, "On the performance of multigroup multicast cell-free massive MIMO," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2642–2645, Dec. 2017.

[27] Y. Zhang, H. Cao, and L. Yang, "Max-min power optimization in multigroup multicast cell-free massive MIMO," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–6.

[28] F. Tan, P. Wu, Y.-C. Wu, and M. Xia, "Energy-efficient non-orthogonal multicast and unicast transmission of cell-free massive MIMO systems with SWIPT," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 949–968, Apr. 2021.

[29] M. Zhou, Y. Zhang, X. Qiao, M. Xie, L. Yang, and H. Zhu, "Multigroup multicast downlink cell-free massive MIMO systems with multiantenna users and low-resolution ADCs/DACs," *IEEE Syst. J.*, vol. 16, no. 3, pp. 3578–3589, Sep. 2022.

[30] M. Farooq, M. Juntti, and L.-N. Tran, "Power control for multigroup multicast cell-free massive MIMO downlink: Invited paper," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 930–934.

[31] H. Lee, S. Moon, Y. Lee, J. Oh, J. Chung, and J. Choi, "Multi-group multicasting systems using multiple RISs," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 9488–9501, Aug. 2024.

[32] G. Femenias, F. Riera-Palou, and E. Björnson, "Another twist to the scalability in cell-free massive MIMO networks," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6793–6804, Nov. 2023.

[33] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends® Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, 2021.

[34] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends® Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017. [Online]. Available: http://dx.doi.org/10.1561/2000000093

[35] J. Li, Q. Pan, Z. Wu, P. Zhu, D. Wang, and X. You, "Spectral efficiency of unicast and multigroup multicast transmission in cell-free distributed massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 12826–12839, Dec. 2022.

[36] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[37] E. Björnson, L. Sanguinetti, and M. Debbah, "Massive MIMO with imperfect channel covariance information," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 974–978.

[38] S. Haghighatshoar and G. Caire, "Massive MIMO pilot decontamination and channel interpolation via wideband sparse channel estimation," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8316–8332, Dec. 2017.

[39] D. Neumann, M. Joham, and W. Utschick, "Covariance matrix estimation in massive MIMO," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 863–867, Apr. 2018.

[40] L. Sanguinetti, E. Björnson, and J. Hoydis, "Toward massive MIMO 2.0: Understanding spatial correlation, interference suppression, and pilot contamination," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 232–257, Jan. 2020.

[41] K. Upadhya and S. A. Vorobyov, "Covariance matrix estimation for massive MIMO," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 546–550, Apr. 2018.

[42] F. Riera-Palou, G. Femenias, A. G. Armada, and A. Pérez-Neira, "Clustered cell-free massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.

[43] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[44] Z. Chen and E. Björnson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018.

[45] A. Á. Polegre, F. Riera-Palou, G. Femenias, and A. G. Armada, "Channel hardening in cell-free and user-centric massive MIMO networks with spatially correlated Ricean fading," *IEEE Access*, vol. 8, pp. 139827–139845, 2020.

[46] R. Nikbakht, R. Mosayebi, and A. Lozano, "Uplink fractional power control and downlink power allocation for cell-free networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 774–777, Jun. 2020.